Negative affectivity and social inhibition in cardiovascular disease

Emons, W.H.M.; Meijer, R.R.; Denollet, J.

Published in:
Journal of Psychosomatic Research

Publication date:
2007

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

Journal of Psychosomatic Research 63 (2007) 27–39

Journal of Psychosomatic Research

Negative affectivity and social inhibition in cardiovascular disease: Evaluating type-D personality and its assessment using item response theory

Wilco H.M. Emons[a,*], Rob R. Meijer[b], Johan Denollet[c]

[a]Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, Tilburg University, Tilburg, The Netherlands
[b]Department of Research Methodology, Measurement and Data Analysis, University of Twente, Enschede, The Netherlands
[c]Department of Medical Psychology, Tilburg University, Tilburg, The Netherlands

Received 22 June 2006; received in revised form 19 February 2007; accepted 1 March 2007

Abstract

Objective: Individuals with increased levels of both negative affectivity (NA) and social inhibition (SI)—referred to as type-D personality—are at increased risk of adverse cardiac events. We used item response theory (IRT) to evaluate NA, SI, and type-D personality as measured by the DS14. The objectives of this study were (a) to evaluate the relative contribution of individual items to the measurement precision at the cutoff to distinguish type-D from non-type-D personality and (b) to investigate the comparability of NA, SI, and type-D constructs across the general population and clinical populations. **Methods:** Data from representative samples including 1316 respondents from the general population, 427 respondents diagnosed with coronary heart disease, and 732 persons suffering from hypertension were analyzed using the graded response IRT model. **Results:** In Study 1, the information functions obtained in the IRT analysis showed that (a) all items had highest measurement precision around the cutoff and (b) items are most informative at the higher end of the scale. In Study 2, the IRT analysis showed that measurements were fairly comparable across the general population and clinical populations. **Conclusions:** The DS14 adequately measures NA and SI, with highest reliability in the trait range around the cutoff. The DS14 is a valid instrument to assess and compare type-D personality across clinical groups.

Keywords: Item response theory; Measurement equivalence; Negative affectivity; Social inhibition; Type-D personality

Introduction

Early identification of cardiovascular patients who are characterized by an unfavorable clustering of psychological risk factors [1] is important in order to improve their prognosis and quality of life. A recent report of the National Heart, Lung, and Blood Institute working group on outcomes research in cardiovascular disease also recommended studies to identify the key determinants of patient-centered outcomes such as quality of life and functional status [2].

In recent years, we have argued that the personality traits of Negative Affectivity (NA) and Social Inhibition (SI) are of special interest in this context [3]. NA denotes the stable tendency to experience negative emotions [4]; high-NA individuals experience more feelings of dysphoria, anxious apprehension, and irritability across time and situations. SI denotes the stable tendency to inhibit the expression of emotions and behaviors in social interaction [5]; high-SI individuals tend to feel inhibited, tense, and insecure when with others. Individuals who are characterized by high NA

Abbreviations: CHD, coronary heart disease; CTT, classical test theory; DIF, differential item functioning; GRM, graded response model; IRF, item response function; IRT, item response theory; NA, negative affectivity; ORC, option response curve; SI, social inhibition.

* Corresponding author. Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, Tilburg University, P.O. Box 90153, 5000LE Tilburg, The Netherlands. Tel.: +31 13 466 2397; fax: +31 13 466 3002.

E-mail address: w.h.m.emons@uvt.nl (W.H.M. Emons).

0022-3999/07/$ – see front matter
doi:10.1016/j.jpsychores.2007.03.010

as well as SI seem to scan the world for signs of impending trouble [6] and avoid negative reactions from others through excessive control over self-expression [7].

Relatively high scores on both NA and SI define the distressed personality type or type-D personality [8]. This type-D personality profile is independently associated with an unfavorable clinical course and poor patient-centered outcomes in various cardiovascular populations, including those with ischemic heart disease [3,8], drug-eluting stenting [9], cardiac arrhythmias [10], peripheral arterial disease [11], and heart failure [12]. The DS14 [13] is a brief self-report measure that was specifically designed for standard assessment of a propensity towards general emotional distress of type-D individuals. The DS14 contains revised items from its predecessor, the DS16 [14], and some new items. The DS14 comprises seven items measuring NA and seven items measuring SI. The content of the items and their underlying lower level constructs of the DS14 can be found in Table 1. A score at or above 10 (range 0–28) on both the NA and SI subscales of the DS14 designates those who have a type-D personality [15]. These choices for the cutoffs were based on the median split in representative samples. Clinical evidence for this cutoff-based type-D classification was obtained in longitudinal clinical studies and empirical evidence from latent class cluster analysis [15].

Despite the apparent promise of the DS14 assessment of NA, SI, and type-D personality in cardiovascular patients, a number of substantive and measurement issues still require further examination. First, from earlier studies it is unclear to what extent the items contribute to reliable classification of type-D and non-type-D individuals using a cutoff of 10 on both the NA and SI scales; that is, more information is needed to document the relative contribution of individual items to the measurement precision of the scale and the reliability of NA and SI assessment around the cutoffs. Items that have the highest relative contribution are the strongest markers of the underlying type-D concept. Second, it is unclear whether there is a difference in item responses between individuals with the same trait values belonging to different clinical populations; that is, individuals surviving an acute coronary event, high-risk individuals without acute coronary event, and individuals from the population at large. Assessment of the comparability across populations is an important part of the validation process when scales are used in different populations. Differences in test and item characteristics between populations may point at substantive qualitative differences [16] in distressed type-D personality that need further exploration.

Both research questions can be more adequately addressed using item response theory (IRT) than using classical test theory (CTT). IRT methods have been applied to measure distress and quality of life in the medical context, including the shortening of scales to measure psychopathology in general, medical wards [17], or quality of life in cancer patients [18], and the rating of musculoskeletal pain in rehabilitation patients [19]. There is, however, a paucity of research on personality in the medical context, including the use of IRT methods in this context. In the present paper, we address this issue by applying IRT analyses to the DS14 assessment of NA, SI, and type-D personality in both individuals from the general population and patients with cardiovascular disorder and hypertension. We first explain the principles of IRT and the advantages of IRT to CTT to analyze the DS14. Second, we report the results of IRT analyses with an emphasis on the relative contribution of individual items to the measurement precision. Third, we focus on the comparability of NA, SI, and type-D assessment in qualitatively distinct groups. Finally, we discuss how the DS14 can be improved in future scale revisions.

Item response theory

Psychological variables, such as NA and SI, cannot be observed directly. These psychological variables are referred

Table 1
Item content and lower level construct for the items of the DS14

Item	Content	Position in DS14	Lower level construct
Negative affectivity			
NA1	Worries about unimportant things	2	Anxious apprehension
NA2	Often feels unhappy	4	Dysphoria
NA3	Is easily irritated	5	Irritability
NA4	Takes gloomy view of things	7	Dysphoria
NA5	Is often in a bad mood	9	Irritability
NA6	Often worries about something	12	Anxious apprehension
NA7	Is often down in the dumps	13	Dysphoria
Social inhibition			
SI1	Makes contact easily	1	Social poise (reversed keyed)
SI2	Often talks to strangers	3	Social poise (reversed keyed)
SI3	Inhibited in social interactions	6	Discomfort in social situations
SI4	Difficulties starting a conversation	8	Discomfort in social situations
SI5	Closed kind of person	10	Reticence
SI6	Keeps others at a distance	11	Reticence
SI7	Does not find things to talk about	14	Discomfort in social situations

to as latent traits symbolized by the Greek letter θ. The goal of psychological measurement is to determine a person's position on the latent trait from a set of observed item responses.

The building blocks of IRT are the item response functions (IRFs). These functions describe the relation between the probability of responding in a certain category (e.g., strongly disagree, disagree, agree, strongly agree) and θ, and, thus, link the observed responses to an underlying latent scale. It is usually assumed that the IRF is an increasing function of θ. This means, for example, that when presenting the statement "I often feel unhappy", persons with high NA scores have a higher probability of answering "strongly agree" than persons with low NA scores. The form and location of the IRFs describe the psychometric properties of the items, such as item popularity (the point at the θ scale where there is a 50% change of responding in a certain category) and discrimination power (the magnitude with which the IRF increases across a certain interval on the θ scale).

Most IRT models assume a unidimensional θ and a specified form for the IRF. In this study we used Samejima's [20,21] graded response model (GRM) to investigate the measurement properties of the DS14. The GRM is suitable for analyzing ordered response categories, such as Likert-type rating scales. Several researchers used this model to analyze personality data (e.g., Refs. [22,23]). The items in the GRM are defined by a slope parameter (a) and two or more location parameters (b); the number of location parameters per item is equal to the number of response categories minus 1. The magnitude of the slope parameter reflects the degree to which the item is related to the underlying latent trait. This means that for high a-values the response categories accurately differentiate among trait levels. The location parameters reflect the spacing of the ordered response categories along the θ-scale. The location parameter b for category m can be interpreted as the point at the latent scale where there is a 50% change of scoring in category m or higher. Thus, respondents with a θ-value higher than b have more than 50% change of responding in category m or higher. These probabilities can be used to determine the option response curves (ORC), which describe the probability of responding in a particular response category conditional on θ. Fig. 1 gives an example of the ORCs for an item with low a-level (upper panel) and high a-level (lower panel). Moving from the lower to the higher end of the θ scale shows that first the zero-category is most likely (low θ levels), then the one-category (medium θ levels), followed by the two-category, and, finally, the three-category (high θ level). Furthermore, the middle category options are more peaked for higher a-values.

Advantages of IRT Analyses

Denollet [13] analyzed the psychometric properties of the DS14 using CTT. Results showed high internal consistency (Cronbach's α >.86) and adequate construct validity. Although CTT provides important group-based information of the reliability and validity of the DS14 scores on the NA and SI scale, it has some limitations in analyzing the diagnostic and screening properties of the scale. First, CTT estimates to what extent observed score differences reflect true-score differences on the underlying psychological trait, but it cannot adequately describe reliability (measurement precision) at specific ranges on the latent trait. An important advantage of IRT over classical theory is that it allows us to determine the test and item characteristics conditionally on θ. Measurement precision in IRT is thus not a constant as in CTT. IRT analysis is more informative about item characteristics than conventional CTT statistics such as item-rest score correlations; that is, IRT tells us whether reliable distinctions can be made at specific points of the scale, both on the item level and test level. Of importance in the present study is that IRT allows assessment of measurement precision at the clinical cutoffs that distinguish type-D and non-type-D personality. The higher the precision at the cutoff, the more reliable individuals can be categorized into different categories. Items that have high precision are strong markers of the underlying type-D concept.

Second, a CTT analysis only demonstrates the item and test characteristics for the population represented by the

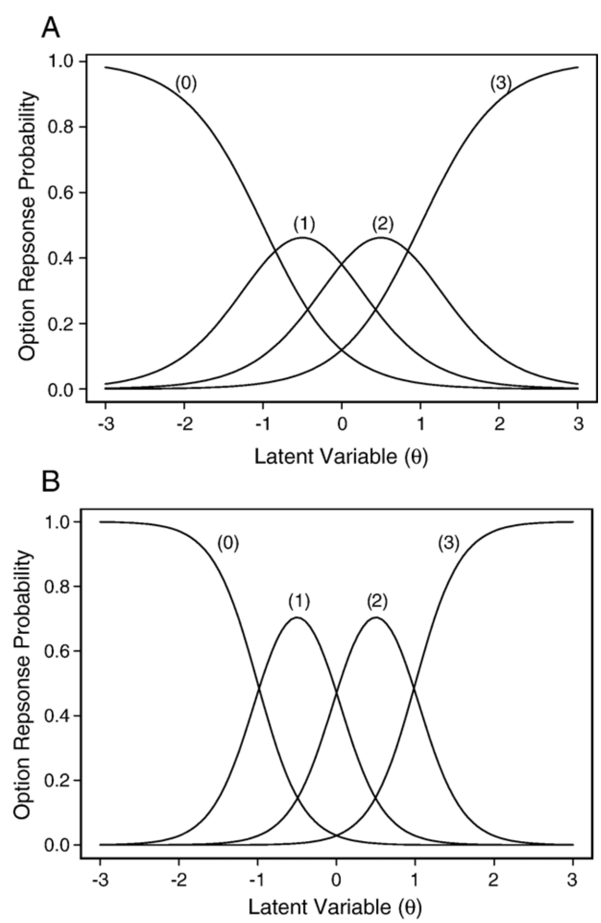

Fig. 1. Option response curves for an item with four response categories, with (A) low discrimination and (B) high discrimination.

specific sample at hand, but these results cannot be generalized to other populations. This limits the usefulness of CTT when analyzing questionnaire data from qualitatively different populations, where groups differ on the underlying construct. To examine carefully the substantive and psychometric properties of the scale, one has to separate the effects of the items, individual differences in the underlying construct, and systematic group differences. In IRT, the relation between test performance and the underlying trait is specified by a mathematical model that separates these three effects. The IRT model can be tested against the data, and once an IRT model is found that describes the data well, one can use that model to describe the test and item characteristics independent of a specific population and vice versa.

The present research

The present research uses the advantages of IRT (1) to further elaborate our understanding of type-D personality and (2) its assessment in qualitatively different groups. Furthermore, we investigated whether the DS14 allows more fine-grid distinctions between levels of type-D personality, and to what extent further scale revisions (e.g., shortening the scale, or reducing the number of item categories) are justified.

(1) The first issue concerns measurement precision and the relative contribution of individual items to the reliable measurement of the underlying construct around the cutoff. The NA scale and SI scale consist of different lower level constructs. As can be seen in Table 1, three out of the seven NA items of the DS14 are related to dysphoria (items NA2, NA4, and NA7), two items are related to anxious apprehension (items NA1 and NA6), and two items measure irritability (items NA3 and NA5). We use the term “dysphoria” here because Kendall et al. [24] have argued that individuals reporting increased levels of depressive symptoms should be referred to as “dysphoric” and that the term “depression” should be reserved for individuals with a clinical diagnosis of affective disorder. Three out of the seven SI items measure discomfort in social situations (items SI3, SI4, and SI7), two items measure lack of social poise (items SI1 and SI2), and two items measure reticence (items SI5 and SI6).

Measurement precision conditionally on θ is assessed by means of the information curves. These curves can be derived for each item and added together to form the test information curve (e.g., Ref. [25]). Because the total information is the sum of the item information functions, a clear picture is obtained of the relative contribution of lower level constructs and their constituent items to the meaning of the type-D construct. In addition, measurement precision defined on the θ-scale can be related to the total score (X_+) through the option response curves. Reliability of classification based on X_+ can be examined locally on the θ-scale and independent from the population of individuals. Information functions can also be used to examine the information provided by different item categories in distinguishing type-D from non-type-D. With the use of these properties, a more concise measurement instrument can be constructed with a minimal loss of measurement precision, as long as the remaining items provide enough information for reliable individual decision making for the application envisaged.

Research showed that reliability measures from CTT, such as Cronbach’s α, are inappropriate to determine whether we can reliably classify persons into different meaningful score categories [25,26]. A good example was recently provided by Langenbucher et al. [27] who considered the role of IRT methods in examining the basis for diagnostic criteria used for classifications according to the *Diagnostic and Statistical Manual of Mental Disorders* (fourth edition) [28]. Their IRT analyses of scale properties of a popular measure of substance abuse (alcohol, cannabis, and cocaine criteria) identified the lack of reliable measurement at the cutoffs that define different *DSM-IV* criteria. By analogy, this raises the question whether the scale properties of the DS14 are useful to classify patients into different categories and whether scale revisions are appropriate.

(2) The DS14 is used in different clinical populations. The interpretation of the differences in scores of individuals and groups from different populations depends on the invariance of the measurement model across the populations. If invariance does not hold, then the test is biased against a particular population and comparisons of individuals across populations may be invalid. Therefore, it is important to establish empirically whether there is a difference in item responses between members with the same trait values belonging to different populations. If the probability of a correct response differs between groups, an item is subjected to differential item functioning (DIF). DIF is thus a way to establish measurement invariance. Establishing measurement invariance is important to progress in many domains of psychological research [29]. In the context of personality assessment, IRT-based DIF analyses were conducted by, for example, Smith and Reise [30] who studied gender differences on measures of NA and neuroticism.

Study 1: IRT item and scale analysis

Method

Participants and measures

Part of the data from Denollet’s [13] study were used. In particular, we used samples that were matched by age range, resulting in data from the Dutch and Belgian population consisting of 2475 (1497 males and 978 females) respondents whose ages ranged from 40 through 87 (mean=57.73 and S.D.=9.80 for males and mean=56.36 with S.D.=10.41 for females). The sample included 1316 (720 males, 596 females) persons from the general population, 427 (399 males, 28 females) respondents diagnosed with coronary heart disease (CHD), and 732 (378 males, 354 females)

Table 2
Estimated item parameters (standard error) of the graded response model for the negative affectivity and social inhibition items

	Slope parameter	Location parameters			
Item	*a*	b_1	b_2	b_3	b_4
Negative affectivity					
NA1	1.47 (0.06)	−1.30 (0.06)	−0.45 (0.04)	0.45 (0.05)	1.70 (0.08)
NA2	2.36 (0.08)	−0.25 (0.03)	0.51 (0.04)	1.29 (0.05)	2.15 (0.09)
NA3	1.45 (0.06)	−1.37 (0.06)	−0.39 (0.04)	0.81 (0.06)	2.18 (0.11)
NA4	2.76 (0.09)	−0.43 (0.03)	0.32 (0.03)	1.16 (0.04)	2.05 (0.07)
NA5	1.80 (0.07)	−0.43 (0.04)	0.55 (0.04)	1.81 (0.08)	2.70 (0.13)
NA6	1.93 (0.07)	−1.58 (0.05)	−0.70 (0.04)	0.16 (0.04)	1.41 (0.06)
NA7	3.61 (0.13)	−0.30 (0.03)	0.38 (0.03)	1.10 (0.04)	1.78 (0.06)
Social inhibition					
SI1	2.49 (0.08)	−0.90 (0.04)	0.07 (0.03)	0.99 (0.04)	1.85 (0.06)
SI2	1.33 (0.06)	−0.91 (0.08)	−0.43 (0.05)	0.70 (0.06)	1.64 (0.09)
SI3	2.14 (0.08)	−0.54 (0.04)	0.18 (0.03)	1.14 (0.05)	2.26 (0.09)
SI4	2.69 (0.09)	−0.44 (0.03)	0.24 (0.03)	0.91 (0.04)	1.82 (0.06)
SI5	2.05 (0.07)	−0.63 (0.04)	0.07 (0.03)	0.86 (0.04)	1.92 (0.07)
SI6	1.52 (0.06)	−1.03 (0.06)	−0.14 (0.04)	1.01 (0.06)	2.53 (0.12)
SI7	2.19 (0.08)	−0.60 (0.04)	0.24 (0.03)	1.32 (0.05)	2.29 (0.09)

Cutoff at θ=0.000.

respondents suffering from hypertension. This sample was used in Study 1 and in Study 2.

As discussed above, the DS14 is a brief questionnaire that measures NA and SI. Subjects rate their personality profile on 5-point Likert scales (0=*false*, 1=*rather false*, 2=*neutral*, 3=*rather true*, and 4=*true*). The sample from the general population had a mean score of 6.50 (S.D.=5.46) for NA and 9.75 (S.D.=6.45) for SI; the CHD sample had a mean score of 8.82 (S.D.=6.34) for NA and 9.59 (S.D.=6.54) for SI; and the hypertension sample had a mean score of 12.84 (S.D.=5.85) for NA and 12.07 (S.D.=5.86) for SI.

Analyses

IRT analyses were conducted using MULTILOG 7 [31] and MULTILOG program default options were used, except for the number of iterations in the estimation procedure. This option was set to 75 to ensure that the parameters were estimated with enough precision. The models were estimated such that the zero point on the θ scale corresponds to the cutoff on both scales.[1] Thus, positive θ values indicate type-D personality, and negative θ values non-type-D personality. The utility of the postulated IRT model depends upon the extent to which a model accurately describes the association in the data. Therefore, for a valid interpretation of the item characteristics, the data have to be consistent with the assumptions underlying the applied IRT model. Determining the dimensionality is a critical issue in IRT, and the correct number of latent factors must be identified a priori. A common procedure to determine the dimensionality of sets of items is by means of factor analysis (e.g., see Refs. [25,37]). Earlier research [13] revealed two dominant traits and showed that all of the NA and SI items loaded between 0.62 and 0.82 on their corresponding factor and between −0.05 and 0.24 on the other factor (with the exception of item SI6, which had a loading of 0.34 on the other factor). Furthermore, both the NA and SI items had high inter-item correlations resulting in a Cronbach's α=.88 for the NA scale and α=.86 for the SI scale, respectively. The correlation between the total scores on the two scales was .37. In addition, we compared the expected and observed IRFs and found good fit of the IRT model.

These results support the assumption of two unidimensional scales, each measuring a different aspect of type-D personality.

Results and discussion

Table 2 shows the estimated *a* parameters and *b* parameters for the NA and SI items. For the NA items, the estimated means of θ-distributions were −0.71 for the general population, −0.38 for the CHD population, and 0.37 for the hypertension population. For the SI items, the estimated means of θ-distributions were −0.14 for the general population, −0.19 for the CHD population, and 0.28 for the hypertension population. Differences between the samples were smaller for SI than for NA. Inspection of the *a*

[1] To account for group differences (i.e., sample from the general population and the two clinical samples), we specified a separate normal θ-distribution for each group, which varied in mean but each with a fixed variance equal to 1. The θ-means were estimated simultaneously with the item parameters. Because in parametric IRT the θ-scale scales are invariant up to linear transformations (i.e., they are interval level scales), one can freely choose (or change) the origin and the unit of the θ scale. One of the practical advantages is that one can choose the origin and unit such that they have a convenient practical interpretation (see Refs. [32–34], for a recent discussion). In this study, the origin of the θ-scales was chosen such that it coincides with the clinical cutoff; that is, θ_{cutoff}=0.000. The unit was specified by fixing the standard deviation of θ in each group equal to 1. Given this origin and unit of the scale, all location parameters can be interpreted as standard normal deviates from the cutoff. This implies that differences between θ values (and between *b*-values for the items) can be interpreted as effect sizes. Goodness-of-fit of the model was investigated using posterior predictive model checks and Monte Carlo simulation [35,36].

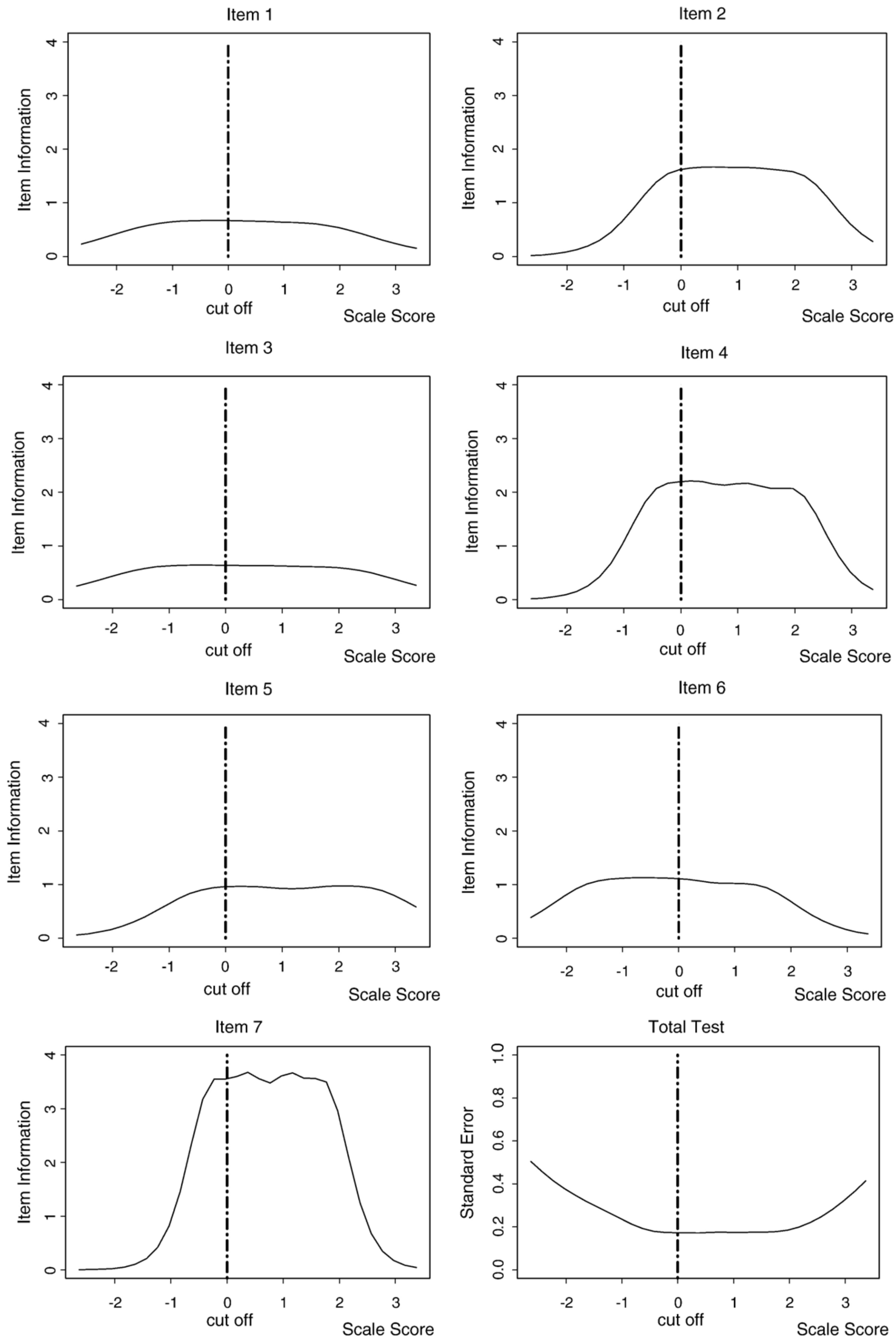

Fig. 2. Item information functions and standard error of the estimated θ the NA items.

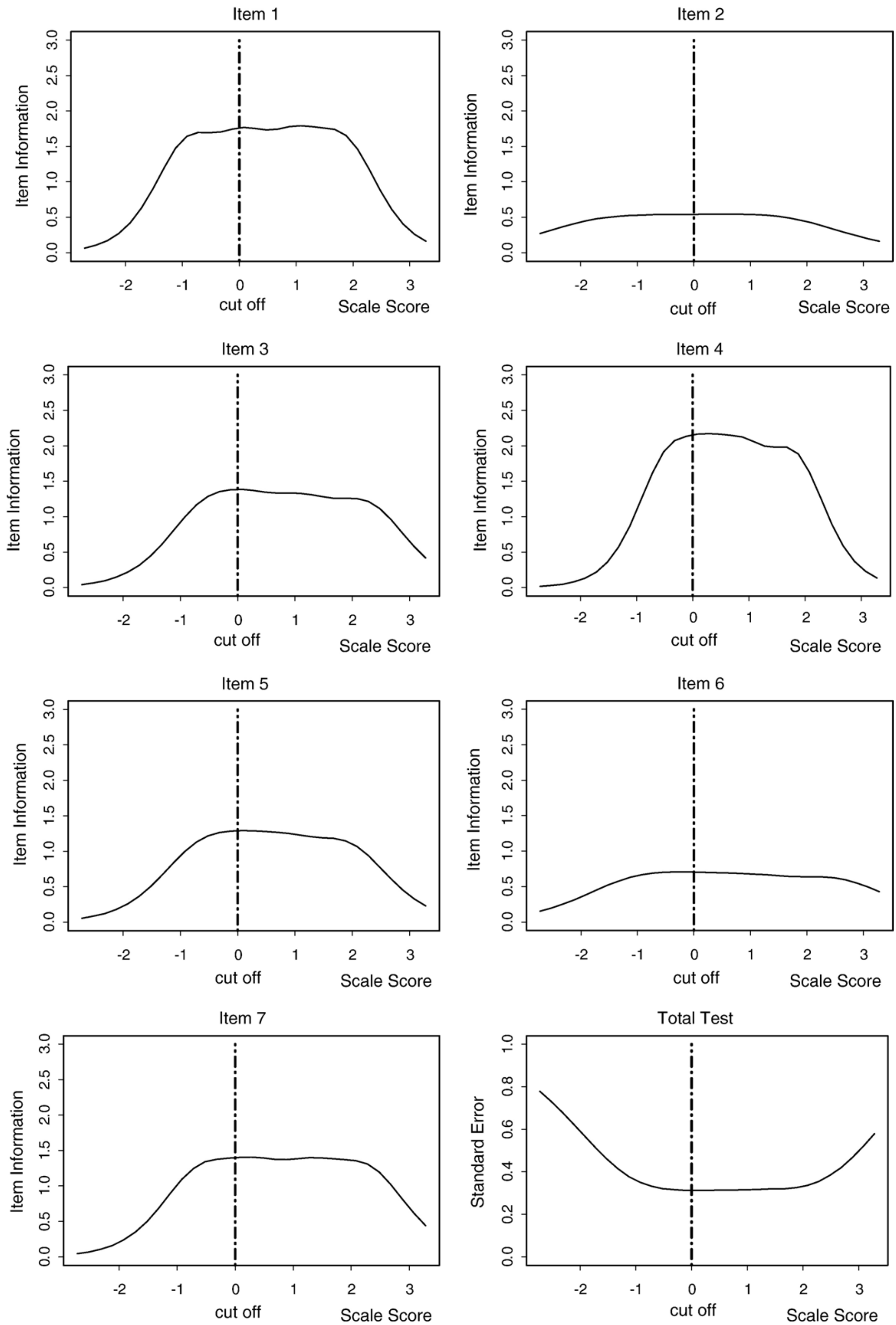

Fig. 3. Item information functions and standard error of the estimated θ for the SI items.

parameters showed differences in the slopes, and inspection of the b parameters showed that the items are located at the higher end of the scale. In particular, for most of the items the higher score categories (2, 3, and 4) only reveal individual differences in NA and SI within the type-D range (i.e., b_3 and $b_4>0$), but they are uninformative about individual differences at the range on the θ-scale where the distinction between type-D and non-type-D is made.

Examining the item information functions around the cutoff of the NA and SI items (Figs. 2 and 3) provides additional information about the importance of lower level indicators. Note that the information functions corroborate the result that the items provide great information at the higher levels along the continuum. For both NA and SI scales, the information curves at the cutoff $\theta=0$ in Figs. 2 and 3 showed that all lower level constructs are indicators of type-D vs. non-type-D personality. The three dysphoria items (items NA2, NA4, and NA7) provide the most information; in particular, items NA4 and NA7 are very distinctive in type-D vs. non-type-D. The three dysphoria items together account for approximately 70% of the information in the NA scores. For SI, the strongest indicator was SI4 "difficulties starting a conversation", followed by SI1 "makes contact easily". Together these indicators provided 40% of the information at the cutoff. The weakest indicators were SI2 "often talks to strangers" and SI6 "keeps others at a distance".

To further study the reliability around the cutoff, it must be noted that in IRT measurement precision is defined on the latent θ scale, whereas classification is based on observed sum scores (X_+). To relate measurement precision on the θ-scale to the X_+ scale, we used a simple Monte Carlo procedure. For each of the θ-levels, $-3.0, -2.9, \ldots, 3.0$, we simulated 1000 item-score vectors under the fitted GRM using the estimated slopes and location parameters (Table 2).[2]

For NA, Fig. 4 shows for each θ-level the mean of the replicated sum scores and the 90% confidence envelopes. The steepness of the function is an indication of the reliability. Note that for persons with $\theta=0$ the mean sum score equals 10, which is implied by the way in which the scale is identified. Fig. 4 shows that for a fixed mean sum score of 10, the corresponding θ scale ranges approximately from -0.5 through 0.5. Persons below the cutoff are measured somewhat less reliably (slope relatively flat) than persons scoring above the cutoff. As a consequence, the predictive validity may be lower for non-type-D individuals due to stronger attenuation effects in this group. Furthermore, Fig. 4 shows that sum scores between 10 and 20 were assessed with approximately equal precision because the slope of the mean-score regression curve is the same within this score region. Similar results were found for the SI scales (graph not presented). Finally, Fig. 4 shows the consistency with which individuals are classified [38]. Persons in the vicinity of the cutoff are consistently classified in the type-D category in about 50% of the cases. This percentage increases when persons are located further away at the right of the cutoff, and it exceeds 90% if the person's θ-value is located more than half a standard deviation from the cutoff.

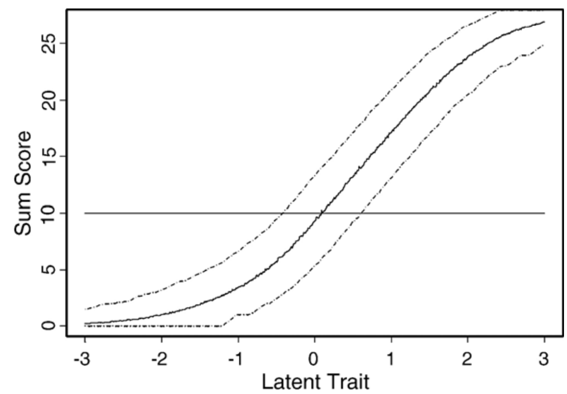

Fig. 4. Nonlinear regression curve of sum score (solid lines) and 90% confidence envelopes (dashed lines) on θ for negative affectivity, based on simulated item-score vectors under the estimated IRT model.

Study 2: Comparability across groups

In Study 2 we focused on the comparability of NA, SI, and type-D assessment in qualitatively distinct groups: persons from the general population, with CHD, and with hypertension. In IRT, measurement invariance is often investigated by means of DIF analysis (e.g., Ref. [39]). An item shows DIF if two respondents having equal levels of θ, but coming from distinct groups, have different probabilities of endorsing each response category of that item. This means that the probability of answering in a particular category depends on group membership. A distinction can be made between full measurement invariance and partial measurement invariance. Full measurement invariance holds when none of the items shows DIF. Partial measurement invariance [40] holds if the majority of the items show no DIF. Although many researchers suggest discarding items showing DIF, the presence of items need not eventuate in biased measurement [29]. Whether DIF items need to be removed needs to be established on a situation-specific basis.

A disadvantage of statistical DIF analysis is that, given enough power, many items in a test may show DIF. However, this DIF may have little practical impact on the assessment at the individual level. One way to address these concerns is by means of person-fit statistics as suggested by Reise and Flannery [41]. Person-fit statistics are developed to identify aberrant response patterns (e.g., Refs. [41–43]). These statistics compare for each person the scores and the

[2] For each simulated vector we calculated the sum score resulting in 1000 replicated X_+ score at each θ level. For each θ-level we calculated the mean sum score and standard deviation across the 1000 replicates, which summarize the distribution of X_+ given θ. This distribution illustrates the influence of random error in the sum score for persons at a particular point θ. One can relate the conditional distribution to the clinical cutoff to evaluate to what extent classifications based on the sum score are inconsistent due to random error [38].

expected item scores under the postulated IRT model. A commonly used IRT-based person-fit statistic is l_z [44], which is the standardized log-likelihood of an individual pattern under the postulated IRT model given a respondent's estimated θ. Large negative values of l_z (say, below -2) indicate that the person's vector is unlikely under the model. Although designed as a measure of person fit, l_z can also be used to assess model fit [41] and to evaluate the comparability of measurements across groups [29].

Method

DIF analysis

For the DIF analysis, we used a two-step procedure. In the first step, the item parameters were estimated separately in each group. If the data fit the GRM and measurement invariance holds, the parameter estimates from different samples are the same up to a linear transformation, even if the samples differ in ability [25]. This means that the item parameters obtained in the sample from the general population, CHD sample, and hypertension sample must be scattered along a straight line. Graphical inspection of plots displaying the estimated parameters may reveal potential DIF items for which the parameter estimates are not linearly related across the samples.

In the second step of DIF analysis, the significance of potential DIF items was tested using a log-likelihood ratio test [25]. First, the log-likelihood is obtained for the model in which the item parameters are constrained to equality in each group (i.e., the null hypothesis of measurement invariance). This model serves as the baseline model. Second, the log-likelihood is obtained for the model in which one or more parameters of potential DIF items are freely estimated within each group. The item parameters for items for which no DIF was found were fixed to be constant across groups to link the scales from different groups. Large differences in log-likelihood between the models indicate significant worse fit of the restricted model compared with model with freely estimated parameters. This indicates that the item contains DIF. Research suggested that DIF tests based on the log-likelihood ratio are more powerful than DIF index approaches (e.g., Ref. [45]). For a technical discussion, the readers are referred to Reise et al. [29].

We used different set-ups for the models with freely estimated parameters, each indicating a different type of DIF. The first model that was tested against the baseline model allows the b parameters of potential DIF item under study to freely vary across groups, but assumes a constant slope parameter (a) across groups. This type of DIF is referred to as uniform DIF (e.g., Ref. [39]). This model allows different endorsement probabilities for each response option across groups. The second model assumes constant b parameters, but allows the slope parameter to vary across groups. This model allows differences across groups in the strength of the relationship between the items and the underlying trait. The third model allows both the location parameters and slope parameter to vary across groups. Thus, both differences in strength and endorsement probabilities are allowed across groups. The second and third models are examples of nonuniform DIF. Differences in likelihood ratio of the baseline model and the alternative model are the basis for testing the significance of DIF (e.g., Ref. [37]).

Person-fit analysis

Person-fit analysis was done by means of a log-likelihood statistic, denoted as l_z [44]. Like most person-fit statistics, a major problem of l_z is that its empirical sampling distribution does not comply with the theoretical sampling distribution (e.g., Ref. [46]) because an estimated θ is used to compute l_z. In particular, the mean of l_z tends to be positively biased, and the variance tends to be too small [47]. This reduces the interpretability of the person-fit results and reduces the power to detect aberrant response patterns. As an alternative, we used a parametric bootstrap procedure [48] to standardize the likelihood such that it can be interpreted as a standard normal deviate. Preliminary simulations showed that this procedure reduced the bias in the distribution of l_z, resulting in l_z values that were close to a standard normal distribution under the null hypothesis. The bootstrapped l_z statistic is denoted by l_z^b.

Results and discussion

Negative affectivity scale

Fig. 5 (upper panel) shows the estimated a parameters in the sample from the general population, in the CHD sample, and in the hypertension sample. The figure reveals one potential DIF item, Item 2 "often feel unhappy". For this item, deviations from a linear trend were found between the estimated slope parameters in the general population and those estimated in the hypertension sample, and also between slope parameters obtained in the CHD sample and in the hypertension sample.

The likelihood-ratio test of uniform DIF yielded G^2=42.16 (df=8) and P<.0000. A test of differences in slopes, keeping the b's fixed, yielded G^2=57.8 (df=2) and P<.0000. For varying slopes and varying thresholds (i.e., the b's), the likelihood-ratio test resulted in G^2=65.8 (df=10) and P<.0000. Varying thresholds in addition to slopes did not result in significant differences, G^2=8 (df=8; P=.43). As a result, the likelihood-ratio tests corroborate the conclusion of DIF on the slopes of Item 2.

The a estimates for Item 2 in each group, with all other parameters fixed, were equal to 2.04 in the sample from the general population, 2.16 in the CHD sample, and 3.33 in the hypertension sample. Thus, Item 2 showed a stronger relation with the underlying trait in the hypertension sample than in the other two samples. For each item pair, the inspection of the plots of the b parameters obtained separately in each sample corroborated the linear relation between the estimates, which indicated no DIF on the category response probabilities.

Research by Smith and Reise [30] revealed DIF between gender groups on measures of neuroticism and NA. To further investigate the role of gender as a possible explanation of the observed DIF of NA2 between the general population and two clinical populations, we repeated our DIF analyses separately in the group of males and females. A test on differences in slopes of NA2 in the group of men, keeping all other things fixed, yielded G^2=14.5 (df=2; P<.0000). The estimated slope under the condition of no DIF was 2.62; allowing the slopes to vary across groups yielded estimated slope parameters of 2.32 in the general sample, 2.36 in the CHD sample, and 4.05 in the hypertension sample. These results corroborated DIF on the slope of NA2. For females, we only compared the normal sample with the hypertension sample because there were only 28 women in the CHD sample. The likelihood ratio test of DIF on the a parameters yielded G^2=11.8 (df=1, P<.000). The estimates were 2.16 in the general sample and 2.88 in the hypertension sample. Although this means that the hypothesis of no DIF was rejected, the differences between the a estimates were smaller for females than for males. Thus, an overall conclusion is that the observed DIF cannot be explained by gender differences alone.

To further explain differences in fit between groups, a regression analysis was done with the dependent variable the l_z^b index and covariates diagnostic group (normal population, CHD, and hypertension), sex, and age (Table 3). For NA (upper panel), only significant effects were found for the hypertension group and sex, but differences were minor. No significant effects were found for age, and no interaction effects were found between sex and diagnostic group.

Table 3
Regression analysis of person-fit results (bootstrapped l_z^b) on age, sex, and diagnostic groups

	Person-fit results			
Indicator	B	SE	P	|effs|[a]
Negative affectivity				
Sex[b]	−0.135	0.045	.003*	0.135
Age	−0.001	0.002	.502	0.047
Diagnostic group[c]				0.258
CHD	−0.071	0.061	.247	
Hyper	0.187	0.051	.000*	
Social inhibition				
Sex[b]	−0.096	0.048	.153	0.096
Age	−0.005	0.002	.056	0.235
Diagnostic group[c]				0.081
CHD	−0.005	0.066	.953	
Hyper	0.076	0.054	.126	

[a] Effs indicates effect size, measured as maximum direct effect.
[b] Reference group is males.
[c] Reference group is the general population.
* Result was significant at the 1% significance level.

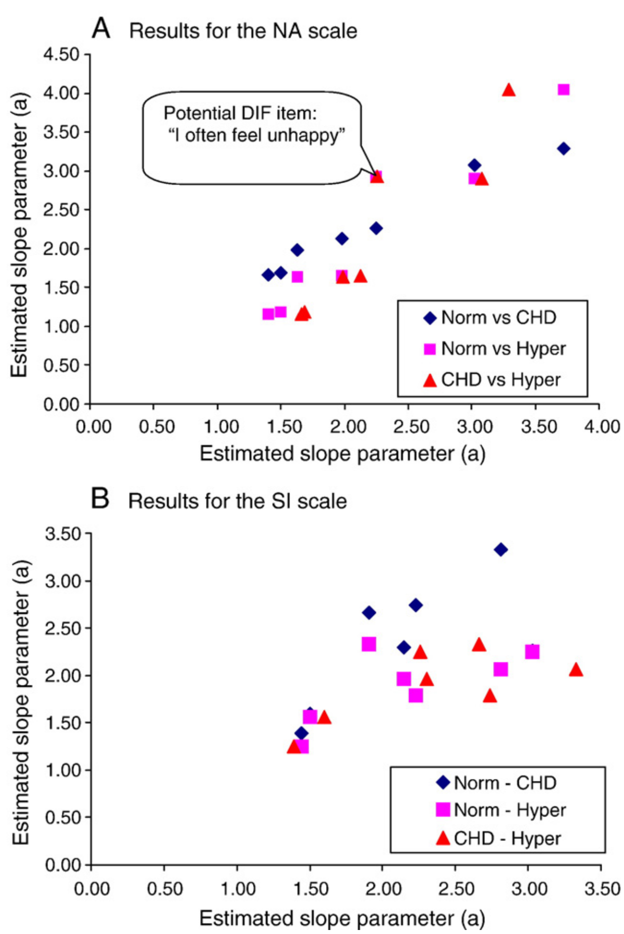

Fig. 5. Across-group comparisons of estimated slope parameters for the NA scale (upper panel) and the SI scale (lower panel). Estimates were obtained separately in the sample from the norm population, and the two clinical samples. Norm=Normal population; CHD=CHD sample; and Hyper=hypertension sample.

Social inhibition scale

Fig. 5 (lower panel) shows the estimated slope parameters for the SI items in the different samples. These estimates show more variation across the samples than the estimates for the NA scale. As a result, it is difficult to draw conclusions about the potential DIF of a particular item. Because the number of items is small, there is a limited amount of information to establish a linear trend and it is difficult to identify unambiguously those items for which the parameters deviate from the linear trend. Likelihood-ratio tests were conducted to study DIF. Each time a particular item was studied the others were used as anchor items. Although several significant results were found, the inspection of the freely estimated parameters generally showed small differences. The lower panel of Table 3 gives the results of regression analysis of the person-fit results for the SI scale. No significant results were found for any of the predictors. From a practical point of view, the scores may be considered comparable across groups.

General discussion

In this study, we investigated the assessment of NA and SI as indicators of type-D personality in the medical context. Persons with a type-D personality profile have a high position on the personality dimensions NA and SI [13]. It has been argued that psychological research needs to explore newer personality constructs such as type-D to see how such constructs may contribute to our understanding of the pathogenic effects of personality on the development and progression of heart disease [49]. Accumulated evidence suggests that type-D personality is associated with increased risk of mortality, morbidity, and impaired quality of life in cardiac patients [50], and that the type-D construct is equally applicable across different nationalities (e.g., [51,52]).

Accordingly, some authors have recommended the use of the DS14 to screen for personality traits that may complicate the clinical course of patients with CHD [53]. The practical importance of the DS14 questionnaire is its screening function of NA, SI, and type-D personality in cardiac patients, because they are more likely to experience levels of emotional distress that have an adverse effect on the treatment and progression of cardiac diseases [3,8–13]. The DS14 may facilitate early diagnosis of potential psychological risk factors that influence cardiac disease, thereby increasing the likelihood of tailored clinical intervention in high-risk patients who are characterized by increased levels of psychological risk factors.

The objectives of this study were twofold. First, we investigated reliability around the cutoff. IRT analyses showed that the DS14 has its highest information around the cutoffs of 10, which warrants the use of this cutoff score to qualify type-D personality and justifies the DS14 as screening instrument for identifying type-D individuals. The IRT analyses further showed that the items of the DS14 are most informative at the higher end of the scale. This is important information, because these are the ranges of interest where the DS14 could be used in future research to make further distinctions between different manifestations of type-D (e.g., medium risk and high risk). Furthermore, the IRT analyses revealed that the higher score categories are uninformative for distinguishing type-D and non-type-D individuals defined by the clinical cutoffs from previous research. The results suggest that for screening purposes categories may be collapsed to increase the efficiency of the DS14 scale. Recent research has shown that reducing the number of answer categories may have a less detrimental effect on the reliability of individual classifications than reducing test length [38]. Furthermore, collapsing the higher score categories does not alter the meaning of the type-D concept.

Second, we investigated type-D assessment across clinical populations using DIF analysis and person-fit analysis. These findings showed that DS14 measurements were comparable across clinical groups and a group from the general population. No important differences were found between the latent trait structure in the sample from the general population and the clinical sample. This means that the measurements are comparable between the groups.

In summary, all items covering the lower level traits of NA and SI yield the highest reliability in the trait range around the cutoffs. This means that all lower level constructs are markers of type-D personality. DIF analysis and person-fit analysis provided further empirical evidence that NA, SI, and type-D are equivalent across clinical groups, which provides further empirical justification for the DS14 as a valid instrument to assess and compare type-D personality across clinical groups. A practical consequence of no DIF is that no different scoring rules are needed for using the scale in different populations.

A limitation of this study was that we analyzed the data using one specific IRT model: the GRM. Although this model is often used to analyze personality data, the question is whether we would have obtained similar results when using another IRT model. Because the GRM fitted the data well, we do not expect that results would have been dramatically different when we had used other models, but future research may use different models to answer this question. Another limitation was that we used person-fit statistics only as a model fit statistic. Differences in person-fit results were used to explain systematic differences between groups. Person-fit statistics are also used for identifying individual item-score vectors resulting from idiosyncratic responses. Inspection of individual item-score vectors that poorly fit the model may alert test users in unexpected response distortions, such as misunderstanding and ambiguities of the wording of the items. Future research may consider these fit statistics when evaluating individual item-score patterns.

In conclusion, the present findings confirm that the DS14 adequately measures the personality traits NA and SI, with high reliability in both traits at the cutoff of 10. This cutoff is accurate in classifying individuals as type-D vs. non-type-D. Furthermore, the measurement of NA, SI, and type-D was comparable across the general population and clinical population. This study provides new evidence for the notion that the DS14 is a valid instrument to assess and compare type-D personality across clinical conditions.

References

[1] Rozanski A, Blumenthal JA, Kaplan J. Impact on the psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. Circulation 1999;99:2192–217.

[2] Krumholz HM, Peterson ED, Ayanian JZ, Chin MH, DeBusk RF, Goldman L, Kiefe CI, Powe NR, Rumsfeld JS, Spertus JA, Weintraub WS. Report of the National Heart, Lung, and Blood Institute working group on outcomes research in cardiovascular disease. Circulation 2005;111:3158–66.

[3] Denollet J, Vaes J, Brutsaert DL. Inadequate response to treatment in coronary heart disease: adverse effects of Type D personality and

younger age on 5-year prognosis and quality of life. Circulation 2000;102:630–5.
[4] Watson D, Pennebaker JW. Health complaints, stress, and distress: exploring the central role of negative affectivity. Psychol Rev 1989;96:234–54.
[5] Asendorpf JB. Social inhibition: a general-developmental perspective. In: Traue HC, Pennebaker HW, editors. Emotion, inhibition, and health. Seattle (Wash): Hogrefe & Huber Publishers, 1993. pp. 80–90.
[6] Bolger N, Zuckerman A. A framework for studying personality in the stress process. J Pers Soc Psychol 1995;68:890–902.
[7] Eisenberg N, Fabes RA, Murphy BC. Relations of shyness and low sociability to regulation and emotionality. J Pers Soc Psychol 1995;68:505–17.
[8] Denollet J, Sys SU, Stroobant N, Rombouts H, Gillebert TC, Brutsaert DL. Personality as independent predictor of long-term mortality in patients with coronary heart disease. Lancet 1996;347:417–21.
[9] Pedersen SS, Lemos PA, van Vooren PR, Liu TK, Daemen J, Erdman RA, Smits PC, Serruys PW, van Dombrug RT. Type D personality predicts death or myocardial infarction after bare metal stent or sirolimus-eluting stent implantation: a Rapamycin-Eluting Stent Evaluated At Rotterdam Cardiology Hospital (RESEARCH) registry sub-study. J Am Coll Cardiol 2004;44:997–1001.
[10] Pedersen SS, van Domburg RT, Theuns DA, Jordaens L, Erdman RA. Type D personality is associated with increased anxiety and depressive symptoms in patients with an implantable cardioverter defibrillator and their partners. Psychosom Med 2004;66:714–9.
[11] Aquarius AE, Denollet J, Hamming JF, De Vries J. Role of disease status and Type D personality in outcomes in patients with peripheral arterial disease. Am J Cardiol 2005;96:996–1001.
[12] Schiffer AA, Pedersen SS, Widdershoven JW, Hendriks EH, Winter JB, Denollet J. The distressed (type D) personality is independently associated with impaired health status and increased depressive symptoms in chronic heart failure. Eur J Cardiovasc Prev Rehabil 2005;12:341–6.
[13] Denollet J. DS14: standard assessment of negative affectivity, social inhibition, and Type D personality. Psychosom Med 2005; 67:89–97.
[14] Denollet J. Personality and coronary heart disease: the Type-D Scale 16 (DS16). Ann Behav Med 1998;20:209–15.
[15] Emons WHM, Denollet J. Latent class cluster analysis of the Type D personality construct in hypertensive patients and healthy controls. J Pers Soc Psychol 2007 [Submitted for publication].
[16] DeBoeck P, Wilson M, Acton S. A conceptual and psychometric framework for distinguishing categories and dimensions. Psychol Rev 2005;112:129–58.
[17] Fink P, Ørnbøl E, Huyse FJ, de Jonge P, Lobo A, Herzog T, Slaets J, Arolt V, Cardoso G, Rigatelli M, Steen Hansen MS. A brief diagnostic screening instrument for mental disturbances in general medical wards. J Psychosom Res 2004;57:17–24.
[18] Bjorner JB, Petersen MA, Groenvold M, Aaronsons N, Ahlner-Elmqvist M, Arraras JI, Bredart A, Fayers P, Jordhoy M, Sprangers M, Watson M, Young T. Use of item response theory to develop a shortened version of the EORTC QLQ-C30 emotional functioning scale. Qual Life Res 2004;13:1683–97.
[19] O'Connor DP. Comparison of two psychometric scaling methods for ratings of acute musculoskeletal pain. Pain 2004;110:488–94.
[20] Samejima F. Estimation of latent trait ability using a response pattern of graded scores. Psychometrika Monograph No 17, 1969. pp 1–100.
[21] Samejima F. Graded response model. In: Van der Linden WJ, Hambleton RK, editors. Handbook of modern item response theory. New York: Springer, 1997. pp. 85–100.
[22] Fraley RC, Waller NG, Brennan KA. An item response theory analysis of self-report measures of adult attachment. J Pers Soc Psychol 2000;78:350–65.
[23] Cooke DJ, Michie C, Hart SD, Hare RD. Evaluating the screening version of the Hare psychopathy checklist-revised (PCL:SV): an item response theory analysis. Psychol Assess 1999;11:3–13.
[24] Kendall PC, Hollon SD, Beck AT, Hammen CL, Ingram RE. Issues and recommendations regarding use of the Beck Depression Inventory. Cogn Ther Res 1987;17:313–24.
[25] Embretson SE, Reise SP. Item response theory for psychologists. Mahwah (NJ): Lawrence Erlbaum, 2000.
[26] Embretson SE. The new rules of measurement. Psychol Assess 1996;8:341–8.
[27] Langenbucher JW, Labouvie E, Martin CS, Sanjuan PM, Bavly L, Kirisci L. An application of item response theory analysis to alcohol, cannabis, and cocaine criteria in DSM-IV. J Abnorm Psychol 2004;113:72–80.
[28] American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed. Washington (DC): American Psychiatric Association, 1994.
[29] Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychol Bull 1993;114:552–66.
[30] Smith LL, Reise SP. Gender differences on Negative Affectivity: an IRT study of differential item functioning on the multidimensional personality questionnaire stress reaction scale. J Pers Soc Psychol 1998;75:1350–62.
[31] Thissen D, Chen WH, Bock RD. MULTILOG (version 7). [computer software]. Lincolnwood (Ill): Scientific Software International, 2003.
[32] Hart B, Jaccard J. Arbitrary metrics in psychology. Am Psychol 2006;61:27–41.
[33] Kazdin AE. Arbitrary metrics: implications for identifying evidence-based treatments. Am Psychol 2006;61:42–9.
[34] Embretson SE. The continued search for nonarbitrary metrics in psychology. Am Psychol 2006;61:50–5.
[35] Sinharay S, Johnson MS, Stern HS. Posterior predictive assessment of item response theory model. Appl Meas Educ 2006;30: 298–321.
[36] Stone CA, Zhang B. Comparing three new approaches for assessing goodness-of-fit in IRT models. J Educ Meas 2003;4:331–52.
[37] Orlando M, Marshall GN. Differential item functioning in a Spanish translation of the PTSD checklist and evaluation of impact. Psychol Assess 2002;14:50–9.
[38] Emons WHM, Sijtsma K, Meijer RR. On the consistency of individual classification using short scales. Psych Methods 2007;12: 105–20.
[39] Holland PW, Wainer H. Differential item functioning. Hillsdale (NJ): Erlbaum, 1993.
[40] Byrne BM, Shavelson RJ, Muthén B. Testing the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. Psychol Bull 1989;105:456–66.
[41] Reise SP, Flannery WP. Assessing person fit on measures of typical performance. Appl Meas Educ 1996;9:9–26.
[42] Meijer RR. Diagnosing item score patterns on a test using IRT based person-fit statistics. Psychol Methods 2003;8:72–87.
[43] Meijer RR, Sijtsma K. Methodology review: evaluating person fit. Appl Psychol Meas 2001;25:107–35.
[44] Drasgow F, Levine MV, Williams EA. Appropriateness measurement with polychotomous item response models and standardized indices. Br J Math Stat Psychol 1985;38:67–86.
[45] Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of the estimated IRT models. In: Holland P, Wainer H, editors. Differential item functioning. Hillsdale (NJ): Erlbaum, 1993. pp. 67–114.
[46] Emons WHM, Meijer RR, Sijtsma K. Comparing simulated and theoretical sampling distributions of the *U3* person-fit statistic. Appl Psychol Meas 2002;26:88–108.
[47] van Krimpen-Stoop EMLA, Meijer RR. Detection of person misfit in computerize adaptive test with polytomous items. Appl Psychol Meas 2002;26:164–80.
[48] Efron B, Tibshirani RJ. Introduction to the bootstrap. New York: Chapman & Hall, 1993.

[49] Habra ME, Linden W, Anderson JC, Weinberg J. Type D personality is related to cardiovascular and neuroendocrine reactivity to acute stress. J Psychosom Res 2003;55:235–45.

[50] Pedersen SS, Denollet J. Type D personality, cardiac events, and impaired quality of life: a review. Eur J Cardiovasc Prev Rehabil 2003;10:241–8.

[51] Grande G, Jordan J, Kümmel M, Struwe C, Schubmann R, Schulze F, Unterberg C, von Känel R, Kudielka BM, Fischer J, Herrmann-Lingen C. Evaluation der deutschen Type D-Skala (DS14) und prävalenz der Typ D-persönlichkeit bei kardiologischen und psychosomatischen patienten sowie gesunden [Evaluation of the German Type D scale (DS14) and prevalence of the Type D personality pattern in cardiological and psychosomatic patients and healthy subjects]. Psychother Psychosom Med Psychol 2004;54:413–22.

[52] Pedersen SS, Denollet J. Validity of the Type D personality construct in Danish post-MI patients and healthy controls. J Psychosom Res 2004;57:265–72.

[53] Albus C, Jordan J, Herrmann-Lingen C. Screening for psychosocial risk factors in patients with coronary heart disease: recommendations for clinical practice. Eur J Cardiovasc Prev Rehabil 2004;11:75–9.