ORIGINAL ARTICLE

Person misfit on the Inventory of Depressive Symptomatology: Low quality self-report or true atypical symptom profile?

Judith M. Conijn^{1,2} | Philip Spinhoven^{2,3} | Rob R. Meijer⁴ | Femke Lamers⁵

¹Research Institute of Child Development and Education, University of Amsterdam, The Netherlands

²Institute of Psychology, Leiden University, Leiden, The Netherlands

³Department of Psychiatry, Leiden University Medical Center, Leiden, The Netherlands

⁴Faculty of Behavioral and Social Sciences, Psychometrics and Statistics, University of Groningen, Groningen, The Netherlands

⁵Department of Psychiatry, EMGO Institute for Health and Care Research, VU University Medical Center/GGZ inGeest, Amsterdam, The Netherlands

Correspondence

Judith M. Conijn, Research Institute of Child Development and Education, University of Amsterdam, The Netherlands.
Email: J.M.Conijn@uva.nl

Funding information

European Union Seventh Framework Program, Grant/Award Number: PCIG12-GA-2012-334065; Geestkracht program of the Netherlands Organization for Health Research and Development (Zon-Mw), Grant/Award Number: 10-000-1002; VU University Medical Center, Leiden University Medical Center, University Medical Center Groningen.

Abstract

Person misfit on a self-report measure refers to a response pattern that is unlikely given a theoretical measurement model. Person misfit may reflect low quality self-report data, for example due to random responding or misunderstanding of items. However, recent research in the context of psychopathology suggests that person misfit may reflect atypical symptom profiles that have implications for diagnosis or treatment. We followed-up on Wanders et al. (*Journal of Affective Disorders*, 180, 36–43, 2015) who investigated person misfit on the Inventory of Depressive Symptomatology (IDS) in the Netherlands Study of Depression and Anxiety (n = 2,981). Our goal was to investigate the extent to which misfit on the IDS reflects low-quality self-report patterns and the extent to which it reflects true atypical symptom profiles. Regression analysis showed that person misfit related more strongly to self-report quality indicators than to variables quantifying theoretically-derived atypical symptom profiles. A data-driven atypical symptom profile explained most variance in person misfit, suggesting that person misfit on the IDS mainly reflects a sample- and questionnaire-specific atypical symptom profile. We concluded that person-fit statistics are useful for detecting IDS scores that may not be valid. Further research is necessary to support the interpretation of person misfit as reflecting a meaningful atypical symptom combination.

KEYWORDS

atypical depression symptoms, careless and random responding, item response theory, person-fit analysis

1 | INTRODUCTION

Research has questioned the suitability of unidimensional models for capturing the complex patterns of depression observed in clinical reality (e.g. Van Loo, de Jonge, Romeijn, Kessler, Schoevers, 2012), while in practice and in research often a unidimensional model is used. For some persons, total scores then may not adequately reflect the underlying variable that is being measured. Recently, person-fit statistics have been proposed for identifying patients with self-reported symptom profiles that do not conform to unidimensional models (e.g. Wanders, Wardenaar, Penninx, Meijer, & De Jonge, 2015; Wardenaar, Wanders, Roest, Meijer, & De Jonge, 2015).

Person-fit statistics are used to detect response patterns that show misfit with respect to a theoretical measurement model, such as an item response theory (IRT) model (Meijer, Niessen, & Tendeiro,

2015). Although the exact definition of person misfit depends on the specific model assumed for the data, person misfit basically identifies an inconsistent and unlikely combination of item scores. For example, a respondent that endorses on a measure of psychopathology the items reflecting severe symptoms (e.g. suicidal ideation) but not any of the milder symptoms (e.g. feeling hopeless or pessimistic) has a misfitting response pattern. Person-fit statistics are sensitive to non-content based invalid responding such as careless responding, as opposed to content-based responding that may lead to extreme high or low total scores (e.g. faking or malingering).

Person-fit statistics were originally developed to detect invalid test scores in cognitive and educational measurement, for example, due to cheating, lack of motivation or scoring errors (Levine & Drasgow, 1982; Meijer & Sijsma, 2001). However, person-fit statistics also have been evaluated and applied in the context of personality

measurement to detect random responding, response styles, faking, and lack of traitedness (e.g. Conijn, Emons & Sijtsma, 2014; Emons, 2008; Ferrando, 2012; LaHuis & Copeland, 2009; Reise & Flannery, 1996; Woods, Oltmanns, & Turkheimer, 2008; Zickar & Drasgow, 1996).

In psychopathology measurement in mental health-care patients, only recently several person-fit studies have been conducted (Conijn, Emons, De Jong, & Sijtsma, 2015; Conrad et al., 2010; Conrad, Conrad, Dennis, Riley, & Funk, 2011; Wanders et al., 2015; Wardenaar et al., 2015). With few exceptions, these studies interpreted misfitting response patterns as reflecting true (i.e. correctly reported) atypical symptom profiles that may have important implications for diagnosis and treatment decisions. For example, Conrad et al. (2010) found that in a sample of persons at intake for drug or alcohol dependence treatment, misfitting response patterns were likely to represent high suicidal ideation combined with overall low symptoms of depression. They concluded that person-fit statistics could be used to screen for atypical suicide risk. Wardenaar et al. (2015) found that person misfit on the Beck Depression Inventory in a sample of myocardial infarct patients reflected an atypical symptom profile characterized by low somatic complaints but other depressive symptoms indicative of clinical levels of depression.

The interpretation of person misfit as an atypical symptom profile is substantially different from the original interpretation of person misfit as signifying an invalid test score due to unmotivated, careless, or biased responding, and is for several reasons not straightforward. First, due to cognitive deficits common to mental illness (Austin, Mitchell, & Goodwin, 2001), mental health-care patients may be particularly prone to concentration and motivation problems during test taking (Cuijpers, Li, Hofmann & Andersson, 2010; Fervaha & Remington, 2013). Consistently, various studies showed that psychological distress is positively related to the likelihood of producing misfitting response patterns on measures of personality and psychopathology (Conijn et al., 2015; Conijn, Emons, Van Assen, Pedersen, & Sijtsma 2013; Reise & Waller, 1993; Wardenaar et al., 2015; Woods et al., 2008). Second, nowadays patients are increasingly often administered large batteries of tests, for example in routine outcome monitoring (De Beurs et al., 2011; De Vries, Meijer, Van Bruggen & Morey, 2016), which may further decrease motivation and concentration. Third, in mental health care, inconsistent and unlikely response patterns have traditionally been regarded as a sign of potentially invalid test results, not as a sign of true atypical symptom profiles. Specifically, frequently used test batteries in mental health care such as the Personality Assessment Inventory (PAI; Morey, 2007) or the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) include validity scales to detect inconsistent response patterns (e.g. Burchett et al., 2016). Similar to person-fit statistics, these scales aim to detect respondents that provide an unlikely combination of item scores.

Explanatory person-fit research, in which a person-fit statistic is regressed on explanatory variables, provides some insight into whether person misfit in a particular sample is predominantly due to inaccurate responding or due to other causes such as true atypical profiles. For example, research showing that person misfit relates to conscientiousness, education level, verbal skills, validity indices, and response styles (Conijn et al., 2013; Conijn et al., 2014; Conijn, Sijtsma, & Emons, 2016; Ferrando, 2009; LaHuis & Copeland, 2009; Meijer, Egberink, Emons, & Sijtsma, 2008; Schmitt, Chan, Sacco, McFarland, & Jennings,

1999; Woods et al., 2008) suggests that misfitting response patterns may be due to inaccurate responding. However, explained variance in the person-fit statistic was small in these studies and other research shows that person misfit on measures of general psychological distress or depression relates to atypical depression, atypical suicide ideation, melancholic depression, and having an uncommon disorder (Conijn et al., 2015; Conrad et al., 2010; Wanders et al., 2015; Wardenaar et al., 2015). These latter results support person misfit as representing true atypical symptoms: subgroups of patients characterized by distinct, atypical symptoms, may be less likely to provide responses consistent with a questionnaire and corresponding IRT model that is based on a sample of patients with common symptom profiles. However, in these studies, measures of explained variance in person misfit were not always provided. Moreover, no study thus far has simultaneously studied explanatory variables reflecting inaccurate responding and those reflecting atypical symptom profiles. Hence, it is unclear to what extent the misfitting response patterns on psychopathology measures are caused by inaccurate responding (implying invalid protocols that should be discarded) and to what extent they reflect clinically relevant atypical patterns of symptoms (which may have important implications for diagnosis and treatment).

1.1 | Person misfit on the Inventory of Depressive Symptomatology (IDS)

In this study, we investigated the causes of person misfit on the Inventory of Depressive Symptomatology (IDS; Rush, Gullion, Basco, Jarrett, & Trivedi, 1996). We used data ($n = 2,981$) of the Netherlands Study of Depression and Anxiety (NESDA), an ongoing longitudinal cohort study (see Penninx et al., 2008). The majority of participants had either a current or a past depression or anxiety disorder. Since data-collection at each study wave took on average 2.5 to 4 hours per person, we expected lack of motivation and concentration to be a relevant problem in the NESDA.

Wanders et al. (2015) also investigated person misfit on the IDS in the NESDA sample, focusing mainly on the baseline assessment. In the baseline data, they used an exploratory approach with 38 potential predictors to identify clinical predictors of misfit. Based on IDS item scores of the baseline and two-year follow-up (FU) assessment, they concluded that person misfit was primarily due to an atypical symptom profile including high symptoms of mood reactivity and suicide ideation combined with generally mild depressive symptoms. However, the results of Wanders et al. (2015) are inconclusive concerning the nature of misfit because variables related to the quality of self-report patterns were not taken into account. A related limitation is that interpretations of the misfitting response patterns were based on the self-reported IDS symptom profile, not on other, independent assessments or theoretically expected atypical profiles. However, if the atypical symptom combination cannot be confirmed by other assessments or does not relate to established atypical symptom combinations, the atypical response pattern may (partly) be caused by an inaccurate self-report.

Our aim was to assess whether person misfit on the IDS reflects clinically relevant atypical symptom profiles and/or error due to inaccurate responding. To this end, we related person misfit on the

IDS to indicators of self-report quality (e.g. response styles, interviewer and respondent evaluation) and theoretically expected atypical symptom profiles (e.g. atypical depression, melancholic depression, and atypical suicide risk), and compared their explanatory value.

2 | METHODS

2.1 | Participants and procedure

Data came from the ongoing NESDA (Penninx et al., 2008), including five data-collection waves across a time-span of six years. Subjects who could not speak Dutch fluently and subjects with a diagnosis of psychotic, obsessive-compulsive, bipolar, or severe addiction disorder were excluded. The lifetime version of the Composite Interview Diagnostic Interview (CIDI) was used to diagnose depressive and anxiety disorders according to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). The baseline sample ($n = 2,981$) included 1,701 respondents currently diagnosed with depression and/or anxiety disorder, 907 persons at risk of a depression or anxiety disorder (due to having life-time diagnoses of depression, a family history of depression/anxiety, or subthreshold depressive/anxiety symptoms), and 373 healthy respondents. Trained research assistants administered the interview and self-report questionnaires. Participants were paid 15 euros for their participation and reimbursed for their travel costs.

We used data of all NESDA respondents (i.e. not only those with a depression diagnosis) because an atypical depression symptom profile may lead to the absence of a depression diagnosis although substantial depressive symptoms are present. Data analyses were conducted using data of the baseline measurement, the two-year FU, and four-year FU.

2.2 | Measures

2.2.1 | Depression severity

The IDS (Rush et al., 1996) includes 30 items to measure depression symptom occurrence and severity. Respondents are asked to describe themselves with respect to the past seven days. Items are rated on a 4-point scale (0–3) with variable options. Two items addressing either decrease (item 11) or increase in appetite (item 12) were recoded into a single item because only one of the items is answered. The same applies to two items that address increase and decrease in weight (items 13 and 14), respectively. Higher scores indicate a higher severity of symptoms. In the present data, Cronbach's alpha for the total score after recoding the weight and appetite items ranged from 0.94 (baseline) to 0.97 (four-year FU).

2.2.2 | Variables quantifying atypical symptom profiles

2.2.2.1 | DSM-IV depression subtypes

The IDS provides classification of respondents into the DSM-IV depression subtypes of atypical depression (Novick et al., 2005) and melancholic depression (Khan et al., 2006). Respondents were classified as having atypical depression when they had mood reactivity, and at least two of the following symptoms; hyperphagia,

hypersomnia, leaden paralysis, and interpersonal rejection sensitivity. Persons were classified as having melancholic depression when they lacked mood reactivity and reported loss of pleasure in (almost) all activities, in addition to reporting $\geq$ three of the following symptoms: distinct quality of depressed mood, mood is worse in the morning, early morning awakening of at least one hour before usual time, psychomotor retardation or agitation, significant anorexia or weight loss, and excessive or inappropriate guilt. Both subtypes were found to be positively related to misfit on the IDS in the NESDA baseline data (Wanders et al., 2015).

2.2.2.2 | Atypical somatization

Based on results of Leentjens, Verhey, Lijkx & Troost (2000) and Wardenaar et al. (2015), we expected response patterns reflecting atypical somatization: high somatic symptoms combined with low mood/cognition symptoms or the opposite pattern of low scores on somatic items and high scores on mood/cognition items. Atypical somatization was assessed using the Mood and Anxiety Symptoms Questionnaire 30-item (MASQ-30; Wardenaar, Van Veen, Giltay, Penninx, & Zitman, 2010), which includes three 10-item subscales that measure negative affect (NA), positive affect (PA), and somatic arousal (SA). To quantify atypical somatization we subtracted the standardized score on the SA subscale from the standardized total score on the combined NA and PA subscales. Next, we took the absolute value of this difference score. Higher values on this difference score indicated a relative higher deviation between the SA score and the combined PA/NA score.

2.2.2.3 | Atypical suicide risk

Based on Conrad et al. (2010), we expected atypical response patterns representing respondents with high suicidal ideation but few other symptoms of depression. Respondents were classified as having atypical suicide ideation based on their total score on the 5-item screening version of the Scale for Suicidal Ideation (SSI; Beck, Kovacs, & Weissman, 1979) and their standardized total score on the IDS (after excluding item 16 which addresses suicide ideation). Specifically, respondents who had positive suicide ideation on the SSI (a total score > 0 ; Eikelenboom, Smit, Beekman, & Penninx, 2012) combined with a standardized IDS total score 0.5 standard deviations lower than the average total score of respondents with positive suicide ideation (i.e. a SSI total score > 0), were classified as having an atypical suicidal ideation.

2.2.2.4 | Bipolar depression

In NESDA, bipolar symptoms are atypical because patients with a bipolar disorder were excluded from the study. Consistently, Wanders et al. (2015) found a small but positive relationship between misfit and bipolar symptoms in the NESDA baseline data. We used the Mood Disorder Questionnaire (MDQ; Hirschfeld et al., 2000) to assess bipolar symptoms.

2.2.2.5 | Data-driven IDS profile

Based on results of Wanders et al. (2015) in the NESDA baseline and two-year FU data, we computed a "data-driven IDS atypical" variable.

Specifically, the scores on the items “irritable” and “anxious” were subtracted from each of the item scores on “reactivity of mood” and “suicidal ideation”. The four resulting difference scores were summed into a single score for which higher values indicate a more atypical profile. According to Wanders et al. (2015), this data-driven atypical profile is similar to the atypical suicide ideation profile of Conrad et al. (2010).

2.2.3 | Variables related to quality of the self-report

2.2.3.1 | Education level

Education level was divided into three categories based on a participants' highest completed education level: basic (elementary not completed and elementary education), intermediate (lower vocational education, general intermediate education, intermediate vocational education, and general secondary education) and high (higher vocational education, college education, and university education).

2.2.3.2 | Interviewer evaluation

After the interview, the research assistants responded to 17 questions pertaining to their impression of the respondents' data collection. These questions concerned problems during the interview and the completion of the self-report scales such as language problems, requests for help, perceived honesty, tenseness, fatigue, perceived concentration and memory capacity of the respondent. The items either had ordered categorical response options or nominal, but not mutually exclusive, response options. We used categorical principal components analysis (CATPCA; Linting & Van der Kooij, 2012) with optimal scaling in SPSS to summarize the interviewer data into a smaller amount of variables with maximum information from the original variable set. CATPCA showed that the data could be summarized into two relevant dimensions, denoted as “Interviewer psychological-cognitive (psych-cog) evaluation” and “Interviewer language evaluation”. The psych-cog score represented the interviewers general evaluation, perceived concentration and memory problems, and tenseness of the respondent. The language score represented items specifically addressing problems during the completion of the self-report questionnaires and language problems in general. The corresponding component scores were used as predictors of person misfit. The Supporting Information provides a detailed overview of the CATPCA procedure and results.

2.2.3.3 | Respondent evaluation

After completing the interview and questionnaires, participants responded to three questions on either 4- or 5-point rating scales: Question 1 (Q1). How tiring was the research participation for you?, with options ranging from (1) *Completely not tiring* to (5) *Very tiring*; Q2. What is your opinion about the total length of the research?, with options ranging from (1) *It was definitely too long* to (4) *It could have been longer*; Q3. How did you experience participating in the research?, with options ranging from (1) *Enjoyable* to (4) *Annoying*. After recoding Q2, Cronbach's alpha of the total score ranged from 0.55 (baseline) to 0.59 (two-year FU). Higher total scores indicated more negative evaluations.

2.2.3.4 | Response style variables – NEO-Five Factor Inventory (NEO-FFI)

The NEO-Five Factor Inventory (NEO-FFI; Costa & McCrae, 1992) was used to measure response styles because it includes negatively worded items, facilitating response style measurement. The NEO-FFI assesses each of the Big Five traits using a 12-item scale, including each four to seven negatively worded items. Items were rated on a 5-point scale ranging from (0) *strongly disagree* to (4) *strongly agree*. A short 23-item version of the NEO-FFI (assessing only extraversion and neuroticism) was administered on the four-year FU.

Agreement response style (ARS) was quantified as the number of agreements (e.g. score ≥ 3) minus the number of disagreements (e.g. score ≤ 1), divided by the total number of items (Van Herk, Poortinga, & Verhallen, 2004). Extreme response style (ERS) was quantified as the percentage of responses in the most extreme categories (e.g. score 0 or 4) (Van Herk et al., 2004). To optimize response style measurement, we did not use all NEO-FFI item scores in calculating ARS and ERS, but selected balanced (i.e. including an even number of negatively and positively worded items) subsets of items within subscales. This resulted in using a total of 42 item scores on baseline and the two-year FU.* Items having the highest corrected item-total correlation were selected in order to optimize the ARS measurement.

2.3 | Statistical analysis

2.3.1 | Person-fit analysis

Person misfit on the IDS was assessed with respect to the unidimensional graded response model (GRM; Samejima, 1997) using the I_2 person-fit index for polytomous item scores (Drasgow, Levine, & Williams, 1985). Previous research also using the NESDA data, showed the IDS to be sufficiently unidimensional for person-fit analysis based on the I_2 statistic (Wanders et al., 2015). The GRM was estimated using MULTILOG 7 (Thissen, Chen, & Bock, 2003) for each of the three data-collection waves separately. Item parameters were estimated using marginal maximum likelihood estimation and person parameters using maximum a posteriori estimation.

Statistic I_2 was standardized using a bootstrap procedure (De la Torre & Deng, 2008). This procedure generated a person-specific null distribution for each respondent which was used to standardize I_2 . To facilitate interpretation of results, I_2 was recoded such that higher positive values indicated more person misfit, that is, more inconsistent response behavior. For respondents that had an item-score pattern consisting only of the minimum score (i.e. 0-score) or only of the maximum score (e.g. 3-score) we set the I_2 value as a missing value. The I_2 value for these response patterns reflects good person fit, but selecting the same answering category throughout a subscale may be due to a response style, lack of motivation, or exaggerating symptoms (Conijn et al., 2016; Ferrando, 2014; Stukenberg, Brady, & Klinetob, 2000). Hence, including these response patterns in the analysis may lead to non-linear predictor effects because they may reflect a low quality self-report.

Person-fit analysis based on short scales may lead to unreliable person-fit values (Emons, 2008). To gain insight into the reliability of I_2 based on the IDS, split-half reliability estimates were computed for I_2 using the Spearman-Brown formula. To account for potential

differences in test-halves, for each wave we used 50 random splits of IDS items and the average split-half reliability was used as the final reliability estimate.

2.3.2 | Model-fit assessment

To quantify person misfit with respect to the GRM, the IDS data need to satisfy the GRM assumptions of unidimensionality, local independence given the latent trait, and monotone increasing logistic item response functions. We used the Mokken R package (Van der Ark, 2012) to plot the sum score against the mean item score. This procedure allowed a visual test of whether item response functions had a monotone increasing shape. Next, we used confirmatory factor analysis (CFA) for ordered categorical data in Mplus (Muthén & Muthén, 2012) to assess goodness of approximation of the one-factor model (Olino et al., 2012). The CFA model for ordinal data estimates an alternative parameterization of the GRM, the normal ogive version (Forero & Maydeu-Olivares, 2009). CFA instead of IRT procedures were used to evaluate GRM model fit because the CFA framework provides established model-fit criteria, including the root mean square error of approximation (RMSEA) and Comparative Fit Index (CFI).

2.3.3 | Multiple regression analysis

We conducted multiple regression analysis for longitudinal data to predict person misfit from the explanatory variables quantifying either the quality of the self-report or atypical symptom profiles. Specifically, we used Mplus to estimate multiple linear regression models for the data of all three waves simultaneously, while correcting the standard errors of the model parameters for the clustering in the data (Muthén & Muthén, 2012).

Foremost, a baseline model was estimated including only the control variables: IDS total score (e.g. Conijn et al., 2015; Wanders et al., 2015; Wardenaar et al., 2015) and dummy variables representing the data collection waves. Next, extended models were estimated by including either a block of atypical symptom profiles, or a block of explanatory variables related to the quality of the self-report. In the former model, we made a distinction between the data-driven atypical IDS profile and all other theoretically expected atypical profiles, by adding the data-driven profile separately to the model. First, because it may be unfairly biased towards predicting misfit given that it is derived from the NESDA data, and second, to assess the extent that the data driven variable had explained variance in common with theoretical atypical variables. Furthermore, we estimated a model including all explanatory variables.

Several additional analyses were also conducted. First, we tested for interaction effects between ERS and each of the atypical symptom profiles. We expected that high ERS combined with atypical symptom profiles may lead to particularly high misfit because atypical symptom combinations are selected and answered in an extreme way. Second, we assessed the effect of measurement error in the I_z index on our results, by treating I_z as a latent variable with variance fixed to $(1 - I_z \text{ reliability}) \times I_z \text{ sampling variance}$.

In all regression models, missing values were handled by multiple imputation of missing data (10 datasets) using Bayesian analysis (Rubin, 1987; Schafer, 1997) in Mplus. Regression coefficients were tested using two-tailed $\alpha = 0.05$.

3 | RESULTS

3.1 | Missing values

Of the baseline respondents, 2,596 (87%) and 2,402 (81%) also participated at the two- and four-year FU, respectively. The number of respondents with missing person misfit values due to a constant IDS pattern including only 0-scores or 3-scores ranged from 94 (3.2%) on baseline to 151 (6.5%) on the four-year FU. The total number of missing person misfit values (due to either drop-out, constant IDS pattern, or missing IDS scores) ranged from 128 on baseline to 816 on the four-year FU. For respondents with a missing person misfit value on a particular wave, the corresponding data was not included in the data analysis, leaving data of 2,853 (baseline), 2,363 (two-year FU), and 2,165 (four-year FU) respondents.

The percentage of missing values on the explanatory variables varied considerably across waves and variables (see Table 1). The number of cases with complete observations on person misfit and the explanatory variables equaled 2,411 (baseline), 2,049 (two-year FU), and 1,978 (four-year FU). The number of respondents with complete variable values on each of the waves equaled 1,283. Person misfit was positively related to the number of missing variables values on baseline ($r = 0.10$, $p < 0.001$), and the two- and four-year FU ($r = 0.08$, $p < 0.001$), suggesting that those respondents dropping out or producing missing item scores were more likely to show misfit.

3.2 | Sample characteristics

The baseline sample included 2,853 subjects (67% woman) aged 18 to 65 years (mean [M] = 42.0; standard deviation [SD] = 13.1). Most respondents (98%) had Dutch nationality. Respondents' education level was basic (6.8%), intermediate (58.5%), or high (34.7%). The majority of persons (80.0%) had a lifetime history of an anxiety or depression disorder. The baseline sample included 48.9% respondents with a six-month prevalent depression diagnosis (major depressive disorder or dysthymia) and 45.1% respondents with a six-month prevalent anxiety diagnosis. For the two- and four-year FU samples, the distribution of demographic characteristics was practically equal to those of the baseline sample. The number of persons with a depression diagnosis decreased to 32.7% and 25.3% on the two- and four-year FU, respectively. The number of persons with an anxiety diagnosis decreased to 28.4% and 24.2% on the two- and four-year FU, respectively.

3.3 | Model-fit evaluation

Results of the Mokken analyses showed that the IDS included several items that were inappropriate for GRM analyses due to violating the assumption of monotone increasing response functions. The following items were removed based on an (almost) flat item response function: Item 4 ("Sleeping too much"), Item 2 ("Sleep during the night"), Item 9a ("Mood in relation to the time of day"). Next, factor analyses were used to identify potential local dependencies and violations of unidimensionality in the reduced IDS item set. Although the one-factor model showed satisfactory RMSEA and CFI fit values for each of the waves, residual correlations pointed at a problematic local dependency between the combined item 11/12 ("Decreased/increased appetite")

TABLE 1 Descriptive statistics for explanatory variables (baseline) and their correlations with person misfit

Explanatory variable (EV)	Baseline				Two-year FU		Four-year FU	
	N	% (n)	M (SD)	r (EV, misfit)	N	r (EV, misfit)	N	r (EV, misfit)
Control variables								
IDS total score	2848		22.2 (13.9)	.20***	2361	.23**	2164	.24***
Atypical symptom profiles								
Bipolar symptoms (MDQ)	2526	—	3.99 (3.41)	.08***	2363	.06**	2154	.01
Atypical somatization (MASQ)	2528	—	0.71 (0.57)	.12***	2325	.13***	2156	.16***
Atypical suicide ideation (IDS/SSI)	2853	3.4 (98)	—	.03	2355	.02	2160	.04*
Atypical depression DSM-IV (IDS)	2853	5.6 (141)	—	.08***	2320	.13***	2154	.11***
Melancholic depression DSM-IV (IDS)	2853	6.3 (184)	—	.14***	2357	.14***	2161	.16**
NESDA data-driven atypical profile (IDS) ^a	2853	—	-1.99 (2.77)	.20***	2360	.27***	2143	.28***
Quality self-report								
High education level ^b	2853	34.7 (991)	—	-.08***	2363	-.06**	2165	-.07**
Agreement response style (FFI)	2851	—	0.07 (0.16)	.04***	2350	.00	2159	.00
Extreme response style (FFI)	2851	—	0.18 (0.16)	.17***	2350	.15***	2159	.17***
Interviewer psych-cog evaluation	2853	—	0.12 (1.06)	.17***	2363	.15***	2165	.17***
Interviewer language evaluation	2853	—	-0.04 (0.89)	.16***	2363	.09***	2165	.12***
Respondent evaluation	2772	—	6.01 (1.76)	.06***	2267	.09***	2015	.07**

Note: The mean and percentage statistics for the two- and four-year follow-up (FU) are not tabulated since they were practically constant across waves for most variables. The exceptions were the Inventory of Depressive Symptomatology (IDS) mean total score (which dropped to ± 15 on the later waves), the mean Mood Disorder Questionnaire (MDQ) score (which dropped to ± 3.5 on the later waves), and the number of patients classified as having melancholic depression (which dropped to $\pm 3\%$ on the later waves), and the mean agreement response style (ARS) score (which increased to 0.18).

* $p < .05$, ** $p < .01$, *** $p < .001$.

^aWanders et al. (2015).

^bVersus combined middle and low education level.

and the combined item 13/14 ("Decreased/increased weight within the last two weeks"). Of the two items, we removed item 13/14 due to its lower factor loading. The final 24-item dataset showed satisfactory RMSEA (≤ 0.07) and CFI (≥ 0.95) values for the one-factor model for each wave and we concluded that the data was appropriate for person-fit analyses based on the GRM.

Inspection of GRM parameter estimates showed high threshold parameters (> 4.5) for the items assessing "early morning awakening", "restlessness" and "other bodily symptoms", indicating that higher categories on these items were infrequently endorsed. The high discrimination parameters (> 3) of the items assessing "sadness" and "general interest" indicated that they contributed most strongly to the measurement of depression severity. Other five items, including those items corresponding to the data-driven atypical profile (i.e. assessing "suicidal ideation" and "reactivity of mood") scored relatively high on both the threshold and discrimination parameters.

3.4 | Person misfit

The number of respondents classified as misfitting based on $\alpha = 0.10$, ranged from 8.5% (two-year FU) to 9.5% (baseline), and the average person misfit value equaled 0.02 on each wave ($SD \pm 1.23$) and was skewed to the right. Skewness values ranged from 0.95 to 1.21. Correlations between person misfit values across waves ranged from 0.37 (baseline with four-year FU) to 0.44 (baseline with two-year FU). The average split-half reliability coefficient computed for I_z ranged from 0.41 (baseline) to 0.45 (two-year FU).

For the 20 respondents with either highest or lowest person misfit on baseline, we inspected the respondent and interviewer comments on the research participation at each wave (i.e. open questions were administered at the data collection). Seven of the severely misfitting respondents provided comments; two were clearly positive, and three clearly negative (e.g. "headache", "too much repetition"). Interviewer comments reflected for 10 of the misfitting respondents problems possibly interfering with the quality of the self-report (memory problems, possible dementia, fatigue, concentration problems, dyslexia, suspects of over-reporting and somatization, limited insight, and language problems). For three of the misfitting respondents, interviewer comments reflected very depressed mood or – in one case – suicidal ideation. Among the 20 best fitting respondents, four provided positive comments and other four respondents provided critical comments or advice (i.e. too abstract/unclear/repetitive questions). For three of the best fitting respondents, the interviewer comments reflected possible problems with the quality of the self-report scores (concentration problems, tenseness, unfamiliarity with question wording), and for another three respondents comments reflected slowness or indecisiveness in responding.

3.5 | Predicting misfit from atypical profiles and quality of self-report

Table 1 shows for the baseline data the means or percentage scores for the explanatory variables. Categorical explanatory variables were only weakly related. For continuous explanatory variables, the largest correlations (range 0.42 to 0.45) were found between the IDS total score and the following variables: atypical depression, melancholic depression,

respondent evaluation, and the interviewer psych-cogn evaluation. The data-driven atypical IDS profile correlated highest with melancholic depression ($r = 0.21$) and $r \leq 0.10$ with other atypical symptom variables.

First, we assessed for each wave separately how person misfit correlated with the explanatory variables (Table 1). The significant correlations were all in the expected direction and most coefficients were consistent across waves. Next, we estimated multiple regression models for the longitudinal data. Table 2 provides the estimated coefficients and explained variance.

The first model including the control variables ($R^2 = 0.05$), showed a positive effect of IDS symptom severity, and more misfit on both the two-year FU and the four-year FU compared to the baseline measurement. The second model added the atypical classifiers to the model ($R^2 = 0.06$). Results showed the expected positive effects of atypical somatization, atypical suicidal ideation, and melancholic depression, but no effect for the bipolar symptoms or atypical depression. The third model added the data-driven atypical IDS profile to the previous model ($R^2 = 0.15$). Atypical depression and melancholic depression were not significant after accounting for the data-driven profile, but other variables had similar effects as compared to the smaller models. The fourth model ($R^2 = 0.09$) estimated the effects of the variables related to the quality of the self-report, excluding the atypical classifiers. We found the expected negative effects of middle and high-level education on person misfit with respect to basic education level, and the positive effects of ERS and the interviewer evaluation scores. Next, the full model was estimated including all predictors ($R^2 = 0.18$). There were no

substantial differences between the estimated effects in the full model and in Models 3 and 4, suggesting that the atypical variables and quality indicators did not explain common variance in person misfit.

Additional models (not tabulated) were estimated to obtain more detailed results. First, we extended the full model with interaction effects between ERS and atypical symptoms profiles. We found small but significant interaction effects between ERS and the bipolar depression symptoms ($\beta = 0.05$, $p < 0.05$), and ERS and the atypical suicide ideation profile ($\beta = 0.05$, $p < 0.05$). Second, we estimated Models 1 to 5 while correcting for measurement error in the I_2 person-fit statistic. We found that R^2 in the full model increased from 0.16 (uncorrected model) to 0.24 (corrected model). However, the increase in explained variance with respect to Model 1 by adding the block of established atypical symptom profiles ($\Delta R^2 = 0.01$) or the quality indicators ($\Delta R^2 = 0.05$) hardly changed by accounting for measurement error.

4 | DISCUSSION

Our goal was to assess the extent to which person misfit on the IDS reflects atypical symptom profiles, and the extent to which misfit is due to low-quality self-report scores. We provide an extension of the Wanders et al. (2015) study, which also used the NESDA baseline data for regressing person misfit on explanatory variables. Their explanatory variables partly overlap with those in our study (i.e. atypical depression, melancholic depression, bipolar-depression symptoms). The added

TABLE 2 Results of multiple regression analysis predicting person misfit

Explanatory variable	Model 1	Model 2	Model 3	Model 4	Model 5
Control variables					
IDS total score	.23 (.01)***	.18 (.02)***	.29 (.02)***	.18 (.02)***	.25 (.02)***
two-year follow-up	.09 (.02)***	.08 (.02)***	.08 (.02)**	.10 (.02)***	.08 (.02)***
four-year follow-up	.10 (.02)***	.08 (.02)**	.07 (.03)**	.12 (.02)***	.10 (.02)***
Atypical symptom profiles					
Bipolar symptoms (MDQ)		-.03 (.02)	-.01 (.01)		-.01 (.01)
Atypical somatization (MASQ)		.07 (.01)***	.07 (.01)***		.06 (.01)***
Atypical suicide ideation (SSI/IDS)		.18 (.07)*	.08 (.07)		.07 (.07)
Atypical depression DSM-IV (IDS)		.02 (.06)	.00 (.06)		.00 (.06)
Melancholic depression DSM-IV (IDS)		.29 (.08)**	.02 (.08)		-.07 (.08)
NESDA data-driven atypical profile (IDS) ^a			.32 (.02)***		.31 (.02)***
Quality self-report					
High-level education				-.24 (.07)**	-.25 (.07)***
Middle-level education				-.20 (.07)**	-.20 (.07)**
Agreement response style (FFI)				.02 (.01)*	.03 (.01)*
Extreme response style (FFI)				.15 (.02)***	.12 (.01)***
Interviewer psych-cog evaluation				.07 (.02)***	.07 (.02)***
Interviewer language evaluation				.04 (.02)**	.04 (.02)*
Respondent evaluation				-.03 (.02)	-.02 (.02)
R^2	.048	.057	.148	.084	.176
ΔR^2 with respect to Model 1	—	.009	.100	.036	.128

Note: For categorical explanatory variables coefficients are standardized with respect to the dependent variable and for continuous explanatory variables coefficients are standardized with respect to both independent and dependent variables. MDQ, Mood Disorder Questionnaire; MASQ, Mood and Anxiety Symptoms Questionnaire; SSI, Scale for Suicidal Ideation; FFI, Five Factor Inventory; IDS, Inventory Depressive Symptomatology.

* $p < .05$, ** $p < .01$, *** $p < .001$.

^aWanders et al. (2015).

value of our explanatory analysis mainly lies in (1) including variables related to the quality of the self-report, (2) including longitudinal data, and (3) selecting explanatory variables for person misfit based on theoretically expected atypical symptom profiles (while Wanders et al. [2015] used a fully exploratory approach).

Atypical somatization, atypical suicide ideation, and melancholic depression predicted person misfit, but explained variance in person misfit was very small (ΔR^2 1%). The expected effects of bipolar symptoms and atypical depression based on theoretical considerations and the Wanders et al. (2015) study were not confirmed longitudinally. Several indicators of self-report quality explained unique variance in person misfit: education level, response styles, and interviewer evaluation scores. Explained variance by the quality indicators was modest (ΔR^2 4%) but higher than explained variance by the theoretical atypical symptom profiles.

Respondent evaluation may not have had the expected effect on person misfit due to low test-score reliability. Additionally, the result may suggest that person misfit does not relate to low motivation but mostly to lack of insight and low cognitive capability. This finding is consistent with the voluntary nature of the study participation. The relative small effect of ARS compared to the effects of ERS and the psychological-cognitive interviewer score is consistent with simulation research into the performance of the I_2 person-fit statistic. Emons (2008) found that I_2 had higher power for detecting ERS and carelessness/inattention, than for detecting ARS. Next to that, a small ARS effect was expected given that all IDS items are positively worded, and ARS then does not lead to strong response inconsistency.

The data-driven IDS profile outperformed both the theoretical atypical symptom profiles and the self-report quality indicators in terms of explained variance (ΔR^2 9%). A possible explanation for this result is that the likelihood of showing misfit relative to the IRT model depends on the estimated IRT model parameters (Woods et al., 2008): respondents who endorse items having extreme threshold parameters combined with high discrimination parameters, are more likely to obtain strong misfit. However, these parameters are affected by sample properties such as the prevalence of symptoms, and questionnaire properties, such as item formulation and response format (Michaelides, 2010). Consistent with this explanation, the items corresponding to the data-driven IDS profile, "suicide ideation" and "reactivity of mood", had a relative extreme combination of these parameters. Hence, the atypical profiles identified by person misfit may be mainly sample- and questionnaire-specific and only to a small extent reflect established atypical symptom combinations.

The low percentage of explained variance in person misfit is attributable to different factors, including unreliability in the predictors and the outcome variable. Another explanation is the use of general person-fit statistic, thus quantifying all sorts of possible deviations from the IRT model. Variation in I_2 is therefore likely due to many different factors, and different persons may show person misfit due to different reasons. A general index is most practical because no specific hypotheses about the type of misfit need to be specified. However, future explanatory person-fit research may also utilize so-called "optimal" person-fit statistics that test for the likelihood of a specific atypical test behavior (Goegebeur, De Boeck, & Molenberghs, 2010; Levine & Drasgow, 1988).

4.1 | Strengths and limitations

A first strength of this study is that we combined the explanatory variables for person misfit derived from cognitive and personality research with the explanatory variables derived from recent psychopathology research. Second, the longitudinal explanatory analyses were important considering that person misfit was only assessed using 24 items and had low reliability. Third, we related person misfit on the IDS to atypical symptom profiles derived from other self-report and rating scales (MASQ, MDQ, SSI). Some previous research only interpreted the atypical profile by using the self-report scale that was also used to quantify misfit (e.g. Conrad et al., 2010), which may lead to confounding atypical symptom profiles with low-quality self-reports.

We related person misfit to theoretically expected atypical profiles to assess whether I_2 person-fit statistic is sensitive to detect true atypical symptom combinations instead of low quality responding. These "established" atypical profiles were based on previous research in other samples. Underlying our approach were two assumptions. First, the atypical symptom profiles identified in previous research are of substantive interest. Second, inaccurate responding as quantified by the I_2 person-fit statistic, does not "coincidentally" lead to such established atypical symptom profiles. Given that previous research identified subgroups of respondents with the same atypical score profile, and that the I_2 statistic detects non-content based invalid responding (e.g. caused by careless responding or lack of concentration), we expect these assumptions to be tenable.

We aimed to provide an improved version of the person-fit analysis on the IDS, resulting in differences in our data-analytic choices that prevent an optimal comparison of our results with those of Wanders et al. (2015). First, they included only patients with a life-time anxiety or depression diagnosis in their analysis ($n = 2,329$). We included also the 373 "healthy" respondents without a diagnosis in the sample because persons with atypical symptom profiles may not receive a diagnosis. Second, we used a bootstrap method to standardize I_2 person-misfit values based on a person-specific null distribution (De la Torre & Deng, 2008), while Wanders et al. (2015) used the same null-distribution for the complete sample.

Although the NESDA study design allowed us to use a diverse set of explanatory variables, some of these variables were of limited quality for addressing our research question. For example, some predictors had low reliability because they were based on few items, the quality indicators in this study were indirectly measured, and the atypical symptom profiles included as explanatory variables were far from exhaustive. The (small) predictor effects should be interpreted in the context of these limitations. A related limitation is that an exact distinction between misfit due to low quality self-responding and misfit due to severe symptoms was impossible. We found that symptom severity related positively to person misfit, but this relationship may partly be due to the co-occurrence of severe symptoms and motivation and concentrations problems. Consistent with this idea, we found that respondents with higher IDS scores had more negative interviewer and respondent evaluations, suggesting that respondents with more severe symptoms provide lower quality self-report data.

5 | CONCLUSION

The number of applications of person-fit analysis in clinical psychology is increasing (e.g. Conijn et al., 2015; Conrad et al., 2010; Conrad et al., 2011; Wanders et al., 2015; Wardenaar et al., 2015), and recently user-friendly software is made freely available for application of person-fit statistics in non-cognitive measurement (e.g. Ferrando & Lorenzo-Seva, 2016; Meijer et al., 2015). However, application of these statistics and follow-up decisions after identifying a misfitting response pattern require a better understanding of what misfit actually represents in a clinical context.

Our results show that person misfit on the IDS does not strongly relate to established atypical symptom profiles and is partly also due to low quality self-reports (e.g. response styles, memory and concentration problems). The latter result suggests that person misfit is useful as a validity indicator on psychopathology measures, whereas additional research is needed to justify the interpretation of person misfit as reflecting meaningful atypical symptom profiles.

ACKNOWLEDGMENTS

The infrastructure for the NESDA study (www.nesda.nl) has been funded through the Geestkracht program of the Netherlands Organization for Health Research and Development (Zon-Mw, grant number 10-000-1002) and participating universities (VU University Medical Center, Leiden University Medical Center, University Medical Center Groningen). FL has received funding from the European Union Seventh Framework Program (FP7/2007-2013) under grant agreement no. PCIG12-GA-2012-334065.

The authors thank Gunhild Franz for her assistance in preparing the dataset.

Declaration of interest statement

The authors have no competing interests.

ENDNOTE

¹ Considering the small number of items on the four-year FU, all 23 available items were used to compute the ARS and ERS indices. Because the response style variables may be affected by the different FFI scales used across waves, we repeated our analyses using ERS and ARS variables based only on the 23-item version. There were no substantial differences in results.

REFERENCES

- Austin, M., Mitchell, P., & Goodwin, G. M. (2001). Cognitive deficits in depression. Possible implications for functional neuropathology. *British Journal of Psychiatry*, 178(3), 200–206. doi:10.1192/bjp.178.3.200
- Beck, A. T., Kovacs, M., & Weissman, A. (1979). Assessment of suicidal intention: The Scale for Suicide Ideation. *Journal of Consulting and Clinical Psychology*, 47(2), 343–352. doi:10.1037/0022-006X.47.2.343
- Burchett, D., Dragon, W. R., Smith Holbert, A. M., Tarescavage, A. M., Mattson, C. A., Handel, R. W., ... Ben-Porath, Y. S. (2016). "False Feigners": Examining the impact of non-content-based invalid responding on the Minnesota Multiphasic Personality Inventory-2 Restructured Form content-based invalid responding indicators. *Psychological Assessment*, 28(5), 458–470.
- Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *The Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for Administration and Scoring*. Minneapolis, MN: University of Minnesota Press.
- Conijn, J. M., Emons, W. H. M., De Jong, K., & Sijtsma, K. (2015). Detecting and explaining aberrant responding to the Outcome Questionnaire-45. *Assessment*, 22(4), 513–524. doi:10.1177/1073191114560882
- Conijn, J. M., Emons, W. H. M., & Sijtsma, K. (2014). Statistic I_2 based person-fit methods for non-cognitive multiscale measures. *Applied Psychological Measurement*, 38(2), 122–136. doi:10.1177/0146621613497568
- Conijn, J. M., Emons, W. H. M., Van Assen, M. A. L. M., Pedersen, S. S., & Sijtsma, K. (2013). Explanatory, multilevel person-fit analysis of response consistency on the Spielberger State-Trait Anxiety Inventory. *Multivariate Behavioral Research*, 48(5), 692–718. doi:10.1080/00273171.2013.815580
- Conijn, J. M., Sijtsma, K., & Emons, W. H. M. (2016). Identifying person-fit latent classes, and explanation of categorical and continuous person misfit. *Applied Psychological Measurement*, 40(2), 128–141. doi:10.1177/0146621615611164
- Conrad, K. J., Bezruczko, N., Chan, Y. F., Riley, B., Diamond, G., & Dennis, M. L. (2010). Screening for atypical suicide risk with person fit statistics among people presenting to alcohol and other drug treatment. *Drug & Alcohol Dependence*, 106(2–3), 92–100. doi:10.1016/j.drugalcdep.2009.07.023
- Conrad, K. J., Conrad, K. M., Dennis, M. L., Riley, B. B., & Funk, R. (2011). *Validation of the Behavioral Complexity Scale (BCS) to the Rasch Measurement Model*, GAIN Methods Report Version 1.2. Chicago, IL: Chestnut Health Systems http://gaincc.org/_data/files/Posting_Publications/Conrad_et_al_2011_BCS_Rasch_Report.pdf
- Costa, P. T. Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory and NEO Five-Factor Inventory professional manual*. Odessa, FL: Psychological Assessment Resources.
- Cuijpers, P., Li, J., Hofmann, S. G., & Andersson, G. (2010). Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: a meta-analysis. *Clinical Psychology Review*, 30(6), 768–778. doi:10.1016/j.cpr.2010.06.001
- De Beurs, E., den Hollander-Gijsman, M. E., van Rood, Y. R., van der Wee, N. J., Giltay, E. J., van Noorden, M. S., ... Zitman, F. G. (2011). Routine outcome monitoring in the Netherlands: practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical Psychology & Psychotherapy*, 18(1), 1–12. doi:10.1002/cpp.696
- De la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, 45(2), 159–177. doi:10.1111/j.1745-3984.2008.00058.x
- De Vries, R. M., Meijer, R. R., Van Bruggen, V., & Morey, R. M. (2016). Improving the analysis of routine outcome measurement data: What a Bayesian approach can do for you. *International Journal of Methods in Psychiatric Research*, 25(3), 155–167. doi:10.1002/mpr.1496
- Dragow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1), 67–86. doi:10.1111/j.2044-8317.1985.tb00817.x
- Eikelenboom, M., Smit, J. H., Beekman, A. T. F., & Penninx, B. (2012). Do depression and anxiety converge or diverge in their association with suicidality? *Journal of Psychiatric Research*, 46(5), 608–615. doi:10.1016/j.jpsychires.2012.01.025
- Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, 32(3), 224–247. doi:10.1177/0146621607302479
- Ferrando, P. J. (2009). A graded response model for measuring person reliability. *British Journal of Mathematical and Statistical Psychology*, 62(2), 641–662. doi:10.1348/000711008X377745
- Ferrando, P. J. (2012). Assessing inconsistent responding in E and N measures: An application of person-fit analysis in personality. *Personality*

- and Individual Differences, 52(6), 718–722. doi:10.1016/j.paid.2011.12.036
- Ferrando, P. J. (2014). A general approach for assessing person fit and person reliability in typical-response measurement. *Applied Psychological Measurement*, 38(2), 166–183. doi:10.1177/0146621613497532
- Ferrando, P. J., & Lorenzo-Seva, U. (2016). A comprehensive regression-based approach for identifying sources of person misfit in typical-response measures. *Educational and Psychological Measurement*, 76(3), 470–486. doi:10.1177/0013164415594659
- Fervaha, G., & Remington, G. (2013). Invalid responding in questionnaire-based research: Implications for the study of schizotypy. *Psychological Assessment*, 25(4), 1355–1360. doi:10.1037/a0033520
- Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, 14(3), 275–299.
- Goegebeur, Y., De Boeck, P., & Molenberghs, G. (2010). Person fit for test speededness: Normal curvatures, likelihood ratio tests and empirical Bayes estimates. *Methodology*, 6(1), 3–16. doi:10.1027/1614-2241/a000002
- Hirschfeld, R. M., Williams, J. B., Spitzer, R. L., Calabrese, J. R., Flynn, L., Keck, P. E. Jr., ... Zajecka, J. (2000). Development and validation of a screening instrument for bipolar spectrum disorder: The Mood Disorder Questionnaire. *American Journal of Psychiatry*, 157(11), 1873–1875. doi:10.2147/NDT.S67842
- Khan, A. Y., Carrithers, J., Preskorn, S. H., Lear, R., Wisniewski, S. R., Rush, J. A., ... Fava, M. (2006). Clinical and demographic factors associated with DSM-IV melancholic depression. *Annals of Clinical Psychiatry*, 18(2), 91–98. doi:10.1080/10401230600614496
- LaHuis, D. M., & Copeland, D. (2009). Investigating faking using a multilevel logistic regression approach to measuring person fit. *Organizational Research Methods*, 12(2), 296–319. doi:10.1177/1094428107302903
- Leentjens, A. F., Verhey, F. R., Luijckx, G. J., & Troost, J. (2000). The validity of the Beck Depression Inventory as a screening and diagnostic instrument for depression in patients with Parkinson's disease. *Movement Disorders*, 15(6), 1221–1224.
- Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology*, 35(1), 42–56. doi:10.1111/j.2044-8317.1982.tb00640.x
- Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, 53(2), 161–176. doi:10.1007/BF02294130
- Linting, M., & Van der Kooij, A. (2012). Nonlinear principal components analysis with CATPCA: A tutorial. *Journal of Personality Assessment*, 94(1), 12–25. doi:10.1080/00223891.2011.627965
- Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscaleable item score patterns using item response theory: An illustration with Harter's Self-Perception Profile for Children. *Journal of Personality Assessment*, 90(3), 227–238. doi:10.1080/00223890701884921
- Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2015). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment*, 23(1), 52–62. doi:10.1177/1073191115577800
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person-fit. *Applied Psychological Measurement*, 25(2), 107–135. doi:10.1177/01466210122031957
- Michaelides, M. P. (2010). A review of the effects on IRT item parameter estimates with a focus on misbehaving common items in test equating. *Frontiers in Psychology*, 1, 167. doi:10.3389/fpsyg.2010.00167
- Morey, L. C. (2007). *The Personality Assessment Inventory Professional Manual*. Lutz, FL: Psychological Assessment Resources.
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide* (Seventh ed.). Los Angeles, CA: Muthén & Muthén.
- Novick, J. S., Stewart, J. W., Wisniewski, S. R., Cook, I. A., Manev, R., Nierenberg, A. A., ... Rush, A. J. (2005). Clinical and demographic features of atypical depression in outpatients with major depressive disorder: Preliminary findings from STAR*D. *Journal of Clinical Psychiatry*, 66(8), 1002–1011. doi:10.4088/JCP.v66n0807
- Olino, T. M., Yu, L., Klein, D. N., Rohde, P., Seeley, J. R., Pilkonis, P. A., ... Lewinsohn, P. M. (2012). Measuring depression using item response theory: An examination of three measures of depressive symptomatology. *International Journal of Methods in Psychiatric Research*, 21(1), 76–85. doi:10.1002/mpr.1348
- Penninx, B. W. J. H., Beekman, A. T. F., Smit, J. H., Zitman, F. G., Nolen, W. A., Spinhoven, P., ... Van Dyck, R. (2008). The Netherlands study of depression and anxiety (NESDA): Rationale, objectives and methods. *International Journal of Methods in Psychiatric Research*, 17(3), 121–140. doi:10.1002/mpr.256
- Reise, S. P., & Flannery, W. P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education*, 9(1), 9–26. doi:10.1207/s15324818ame0901_3
- Reise, S. P., & Waller, N. G. (1993). Traitendness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, 65(1), 143–151. doi:10.1037/0022-3514.65.1.143
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B., & Trivedi, M. H. (1996). The Inventory of Depressive Symptomatology (IDS): Psychometric properties. *Psychological Medicine*, 26(3), 477–486. doi:10.1017/S0033291700035558
- Samejima, F. (1997). Graded response model. In W. J. van der Linden, & R. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85–100). New York: Springer.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person-fit and effect of person-fit on test validity. *Applied Psychological Measurement*, 23(1), 41–53. doi:10.1177/01466219922031176
- Stukenberg, K., Brady, C., & Klinetob, N. (2000). Use of the MMPI-2's VRIN scale with severely disturbed populations: Consistent responding may be more problematic than inconsistent responding. *Psychological Reports*, 86(1), 3–14. doi:10.2466/pr0.2000.86.1.3
- Thissen, D., Chen, W. H., & Bock, R. D. (2003). *MULTILOG for Windows (Version 7)*. Lincolnwood, IL: Scientific Software International.
- Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, 48(5), 1–27. doi:10.18637/jss.v048.i05
- Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, 35(3), 346–360. doi:10.1177/0022022104264126
- Van Loo, H. M., de Jonge, P., Romeijn, J. W., Kessler, R. C., & Schoevers, R. A. (2012). Data-driven subtypes of major depressive disorder: A systematic review. *BMC Medicine*, 156(10). doi:10.1186/1741-7015-10-156
- Wanders, R. B. K., Wardenaar, K. J., Penninx, B. W. J. H., Meijer, R. R., & De Jonge, P. (2015). Data-driven atypical profiles of depressive symptoms: Identification and validation in a large cohort. *Journal of Affective Disorders*, 180, 36–43. doi:10.1016/j.jad.2015.03.043
- Wardenaar, K. J., Wanders, R. B. K., Roest, A. M., Meijer, R. R., & De Jonge, P. (2015). What does the Beck depression inventory measure in myocardial infarction patients?: A psychometric approach using item response theory and person-fit. *International Journal of Methods in Psychiatric Research*, 24(2), 130–142. doi:10.1002/mpr.1467
- Wardenaar, K. J., Van Veen, T., Giltay, E. J., Penninx, B. W., & Zitman, F. G. (2010). Development and validation of a short version of the mood and anxiety symptoms questionnaire: The MASQ-30. *Psychiatry Research*, 179(1), 101–106. doi:10.1016/j.psychres.2009.03.005

- Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2008). Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychological Assessment*, 20(2), 159–168. doi:10.1037/1040-3590.20.2.159
- Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement*, 20(1), 71–87. doi:10.1177/014662169602000107

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Conijn JM, Spinhoven P, Meijer RR, Lamers F. Person misfit on the Inventory of Depressive Symptomatology: low quality self-report or true atypical symptom profile? *Int J Methods Psychiatr Res*. 2017;26:e1548. <https://doi.org/10.1002/mpr.1548>