Psychological Medicine

cambridge.org/psm

Person-fit feedback on inconsistent symptom reports in clinical depression care

Rob B. K. Wanders[1], Rob R. Meijer[2], Henricus G. Ruhé[3], Sjoerd Sytema[1], Klaas J. Wardenaar[1] and Peter de Jonge[1,4]

[1]Interdisciplinary Center Psychopathology and Emotion regulation (ICPE), University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [2]Department of Psychometrics and Statistics, University of Groningen, Groningen, The Netherlands; [3]Department of Psychiatry, Mood and Anxiety Disorders, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands and [4]Department of Developmental Psychology, University of Groningen, Groningen, The Netherlands

Original Article

Cite this article: Wanders RBK, Meijer RR, Ruhé HG, Sytema S, Wardenaar KJ, de Jonge P (2018). Person-fit feedback on inconsistent symptom reports in clinical depression care. *Psychological Medicine* **48**, 1844–1852. https://doi.org/10.1017/S003329171700335X

Received: 7 March 2017
Revised: 18 October 2017
Accepted: 20 October 2017
First published online: 27 November 2017

Key words:
Decision support techniques; depression; inconsistent symptom reports; person fit; self-report

Author for correspondence:
Rob B. K. Wanders, E-mail: r.b.k.wanders@umcg.nl

Background. Depressive patients can present with complex and different symptom patterns in clinical care. Of these, some may report patterns that are inconsistent with typical patterns of depressive symptoms. This study aimed to evaluate the validity of person-fit statistics to identify inconsistent symptom reports and to assess the clinical usefulness of providing clinicians with person-fit score feedback during depression assessment.

Methods. Inconsistent symptom reports on the Inventory of Depressive Symptomatology Self-Report (IDS-SR) were investigated quantitatively with person-fit statistics for both intake and follow-up measurements in the Groningen University Center of Psychiatry ($n = 2036$). Subsequently, to investigate the causes and clinical usefulness of on-the-fly person-fit alerts, qualitative follow-up assessments were conducted with three psychiatrists about 20 of their patients that were randomly selected.

Results. Inconsistent symptom reports at intake (12.3%) were predominantly characterized by reporting of severe symptoms (e.g. psychomotor slowing) without mild symptoms (e.g. irritability). Person-fit scores at intake and follow-up were positively correlated ($r = 0.45$). Qualitative interviews with psychiatrists resulted in an explanation for the inconsistent response behavior (e.g. complex comorbidity, somatic complaints, and neurological abnormalities) for 19 of 20 patients. Psychiatrists indicated that if provided directly after the assessment, a person-fit alert would have led to new insights in 60%, and be reason for discussion with the patient in 75% of the cases.

Conclusions. Providing clinicians with automated feedback when inconsistent symptom reports occur is informative and can be used to support clinical decision-making.

Introduction

Psychiatrists and psychologists make extensive use of questionnaires in clinical decision-making. The use of item response theory (IRT; Embretson & Reise, 2000) models to construct and to evaluate the psychometric quality of such clinical questionnaires is becoming the standard procedure. The application of these models enables clinical researchers to construct computer adaptive tests, to detect differential item and test functioning, and to assess unexpected test behavior in individuals. Studies focused on this latter topic are grouped under the banner of person-fit research (Meijer & Sijtsma, 2001).

The aim of person-fit research is to detect item-score patterns that are inconsistent compared to the other patterns in the sample, or that are inconsistent given the model that is assumed to describe the data (Meijer, 2003; Ferrando, 2015). Inconsistent response patterns may result in invalid test scores, or at least test scores that are difficult to interpret. For example, unmotivated respondents may more or less randomly fill out a questionnaire, or severely depressed patients may be inconsistent with respect to their true state because they would like to hide certain symptoms (e.g. to prevent unwanted hospitalization). In addition, within psychopathology measurement clinical causes like comorbidity (Wanders *et al.* 2015*a*; Wardenaar *et al.* 2015) or concentration problems (Conijn *et al.* 2016, 2017) may lead to inconsistent symptom reports where patients report severe symptoms without milder symptoms (Woods *et al.* 2008; Conrad *et al.* 2010; Wardenaar *et al.* 2015; Conijn *et al.* 2016). Detecting such inconsistent response patterns may have diagnostic value in clinical practice (Pallant & Tennant, 2007; Reise & Waller, 2009; Thomas, 2011). Furthermore, detecting invalid test scores may have value in routine outcome measurement where test scores are compared across different test occasions and invalid test scores may provide wrong impressions about the patient's stability or change across test administrations.

Although many statistics are available to assess person fit, there are only few studies that examine the reason of this unexpected test behavior and how this behavior can invalidate

CAMBRIDGE UNIVERSITY PRESS

test interpretation by a clinician. Recent person-fit studies have suggested that inconsistent symptom reports could reflect clinical factors, like an atypical suicide risk (Conrad *et al.* 2010), or atypical presentation of symptoms (Wanders *et al.* 2015*a*; Wardenaar *et al.* 2015), but could also reflect causes associated with the quality of self-reported data (Conijn *et al.* 2016), caused by, for example, cognitive problems (Conijn *et al.* 2016, 2017). These studies analyzed large-scale archival data, which makes it difficult to retrieve and study the factors that underlie the responses for specific patients and from a more qualitative point of view (see Meijer *et al.* 2008, for an exception in the education field). Embretson & Reise (in press) noted that *'there is scant evidence that person-fit scores mean anything psychologically about an individual, or are even useful for invalidating scale scores. That is psychometric researchers seem especially adept at creating "new" person fit indices, or "cleaning up" their sampling distributions, but fumble when it comes to studying their validity'*. Indeed, there are many theoretical studies concerning person-fit measurement, but we know of no study that takes the next logical step to incorporate person-fit scores into the assessment process and evaluate its practical merits.

To our knowledge, this is the first study in the clinical field to investigate if there is some psychological reality behind inconsistent or misfitting item-score patterns. That is, we aimed to investigate whether there is a clinically relevant explanation behind inconsistent symptom reports, and if providing clinicians with person-fit feedback would be of clinical use in specialized depression care. The study was conducted in a psychiatric setting, where a computer-based automatic test-administration system was used to assess depression severity in patients that presented to a specialized mental health care institute. Using data from this system, person-fit statistics were calculated for 2036 patients and those with inconsistent response patterns were identified. First, the content validity of these flagged item response patterns was evaluated. Next, to evaluate the possible usefulness of an on-the-fly feedback system for caregivers, we asked three psychiatrists if, in retrospect, an 'inconsistency alert' would have been helpful for them in the interpretation of their patients' depression severity scores.

Material and methods

Participants and procedure

Data came from 2036 patients who completed at least one Inventory of Depressive Symptomatology Self-Report (IDS-SR) questionnaire (see below) in 2014 at the University Center of Psychiatry in Groningen. The IDS-SR was used both to assess depressive symptom severity at intake, and to monitor change. Using an online routine outcome monitoring system (RoQua; http://www.roqua.nl), patients were invited to complete questionnaires before, during, and after treatment. Patients could complete questionnaires at home or at the University Center of Psychiatry, where support was available. Patients were informed that anonymized data could be used for research prior to data collection. Because our study used anonymized data and did not involve treatment interventions, the medical ethics committee of the University Medical Center Groningen waived formal judgement.

After the completion of a questionnaire by the patient, the clinicians obtained a feedback report and could further inspect responses to individual items. Person-fit statistics were implemented within this feedback report (online Supplementary Fig. 1), and data were retrospectively extracted (see below). Whenever a patient completed an IDS-SR assessment, besides the regular feedback an extra alert was visible informing the clinician about possible inconsistencies within the reported IDS-SR symptom pattern based on the person-fit statistic. An extensive explanation of the person-fit alert was given, together with a warning that the alert should be used for research purposes only.

Data for the current study came from the assessment at intake, and from repeated measures during and after treatment. As many patients were still under treatment at the end of 2014, they did not (yet) have a follow-up assessment. During treatment the clinicians are free to reassess their patients, and therefore the number of measurements and the time between measurements varies across patients. Of the 2036 patients, 754 (37.0%) patients were assessed more than once; 273 (13.4%) were measured twice [average time between measurements 127 days (S.D. = 135 days)] and 481 (23.6%) were measured more than two times [average time between measurements 90 days (S.D. = 100 days)], providing a total of 6091 IDS-SR responses. Patient data came from 104 clinicians, and included data on age, gender, and clinical diagnoses. Of the patients in the sample, 1021 (50.1%) completed the IDS-SR at the outpatient clinic for general psychiatry, and 1015 (49.9%) completed the IDS-SR as patients in specialized psychiatric care programs.

As a first pilot to evaluate the validity and potential usefulness of implemented person-fit alerts, we conducted a qualitative follow-up study. The head of the mood disorder clinic of the University Center of Psychiatry was approached to participate in this qualitative follow-up study (H.G.R., coauthor), and he subsequently invited two other psychiatrists to participate. The three psychiatrists were asked in retrospect to give detailed feedback about a total of 20 of their patients (respectively, six, seven, and seven patients per psychiatrist) that were randomly selected from all their patients that were flagged as inconsistent responders. Feedback was obtained through a structured questionnaire (online Supplementary Fig. 2) that contained closed and open questions, both about the nature of the inconsistency as well as the possible usefulness of such an alert for clinical practice. These questionnaires were returned between February and May 2015, with an average time between assessments and qualitative follow-up of 236 days (S.D. = 155 days).

Instrument

The Inventory of Depressive Symptomatology Self-Report (IDS-SR; Rush *et al.* 1996) is a self-report questionnaire consisting of 30 items rated on a 4-point (0–3) anchored scale to assess the severity of depressive symptoms. As a patient could either endorse 'appetite increase' or 'appetite decrease' and either 'weight increase' or 'weight decrease', these items were combined, respectively, into compound 'appetite change' and 'weight change' items. The IDS-SR assesses all DSM-IV criterion symptoms for major depressive disorder (MDD) and the most commonly associated noncriterion symptoms (e.g. anxiety and irritability).

Statistical analysis

Person fit

Person-fit statistics enable the identification of patients for whom the observed symptom pattern is different than would be expected based upon the model used to describe the data (Meijer & Sijtsma, 2001). The model estimates for each symptom how likely it is to be reported at different levels of depression severity. For each

https://doi.org/10.1017/S003329171700335X Published online by Cambridge University Press

patient, it is then expected that milder symptoms are more likely to be reported than more severe symptoms. The more a symptom profile deviates from this pattern, the poorer person fit will be. In many such cases misfit is caused by more severe symptoms being reported (e.g. suicidal ideation) without the reporting of milder symptoms (e.g. sad mood).

In the current study, person-fit analyses were performed using the likelihood-based standardized l_z statistic (Drasgow *et al.* 1985) with the graded response model (GRM; Samejima, 1969) as IRT model describing the data. The GRM was chosen because of the ordinal nature of the IDS-SR response data and describes the relation between symptoms and the underlying depression severity. The l_z person-fit statistic then represents how likely it is to observe the pattern of reported symptoms given this estimated model. Patients with symptom reports that are consistent with the model obtain high values on l_z indicative of good person fit, whereas patients with inconsistent reports obtain low values on l_z indicative of poor person fit. Two parameters are estimated under the GRM to describe each item: the *discrimination parameter* (α) reflects how strong a symptom (item) is related to underlying depression severity, and the *threshold* (β) is reflective of symptom severity. The IRT model was calibrated on a sample of depressed and anxiety patients, who completed the IDS-SR ($n = 2329$; Wanders *et al.* 2015*a*) in the Netherlands Study of Depression and Anxiety (NESDA; Penninx *et al.* 2008).

It was chosen to calibrate the IRT model on a well-defined external sample as the current sample of psychiatric patients consisted of a heterogeneous group with very diverse psychopathology, and fitting an IRT model might result in interpretation problems. The NESDA sample is well defined in terms of psychopathology through the use of a standardized structured CIDI interview, and the use of person-fit statistics was previously investigated in this sample (Wanders *et al.* 2015*a*). Since the purpose of the IDS-SR is to measure depression severity in patients with depression, using the IRT model from the well-defined calibration sample in the person-fit analyses allowed for identification of patients, for whom this model did not hold. Patients with good person fit have symptom patterns consistent with the IRT model, whereas those with poor person fit have symptom patterns that are inconsistent with the model and therefore not a good reflection of depression severity.

Analyses

First, person-fit analyses were performed on the intake IDS-SR assessments of all patients ($n = 2036$). Patients were divided into two groups based on their person fit score compared to a 5% significance level cutoff ($l_z < -1.39$) and to a 1% significance level cutoff ($l_z < -2.21$), obtained from the reference person-fit study in the calibration sample (Wanders *et al.* 2015*a*). Patients with person-fit scores below the cutoff score were allocated to an 'inconsistent group' and were further investigated in terms of symptom patterns and external associations.

Second, stability of person-fit scores across repeated measurements was investigated in patients with follow-up assessments ($n = 754$). Correlation between person fit on first and second measurement and the proportion flagged as inconsistent at each measurement were investigated. In addition, the response patterns of six patients with >18 measurements were randomly selected and investigated in more detail. Here, two patients were selected with no measurements flagged as inconsistent, two with <25% inconsistent, and two with >90% of measurements inconsistent.

Third, results of qualitative follow-up assessments on 20 randomly selected patients with poor person-fit scores of three psychiatrists were investigated. Both explanations of psychiatrists on the potential causes of poor person fit, as well as the potential clinical usefulness of a person-fit alert for psychiatrists at the time of actual measurement, were retrospectively assessed.

All analyses were performed in R using the 'ltm' package for fitting the IRT model (Rizopoulos, 2006) and the 'PerFit' package for person-fit analyses (Tendeiro *et al.* 2016).

Results

The sample had a mean age of 43.6 years (S.D. = 14.5) and included 1061 women (52.1%; Table 1). Patients showed a wide variety of primary clinical diagnoses, with mood disorder (25.1%), and anxiety disorder (15.6%) most prevalent. A secondary clinical diagnosis was observed for 33.6% of the patients with 517 (25.4%) Axis I diagnoses (e.g. anxiety disorder) and 167 (8.2%) Axis II diagnoses (personality disorder).

Person fit at intake

The distribution of person fit at intake (l_z mean = −0.41; S.D. = 1.35) was skewed to the left (Fig. 1), and showed higher person-fit scores for extreme low- and high-depression severity. However, there was no relation between person fit and depression severity ($\rho = 0.03$;95% CI [−0.02 to 0.08]) in those with poor person fit ($l_z < -1.39$).

Of all patients at intake ($n = 2036$), 543 (26.6%) had person-fit scores below the 5% significance level, and 260 (12.8%) below the

Table 1. Descriptives of UCP patient sample ($n = 2036$)

Characteristic	Mean or frequency (S.D. or %)
Male gender	975 (47.9%)
Age	40.8 (15.5)
IDS-SR score	29.9 (15.0)
IDS-SR measurements	3.0 (5.0)
Primary clinical diagnosis	
Anxiety disorder	318 (15.6%)
Bipolar disorder	195 (9.6%)
Childhood and developmental disorder	167 (8.2%)
Mood disorder	512 (25.1%)
MDD – first episode	162 (8.0%)
MDD – recurrent	309 (15.2%)
Other	41 (2.0%)
Personality disorder	120 (5.9%)
Schizofrenia or psychotic disorder	57 (2.8%)
Somatoform disorder	77 (3.8%)
Other clinical disorder	119 (5.8%)
Secondary clinical diagnosis	
Axis I disorder	517 (25.4%)
Axis II disorder	167 (8.2%)

S.D., standard deviation; IDS-SR, inventory of depressive symptomatology-self-report; MDD, major depressive disorder.

https://doi.org/10.1017/S003329171700335X Published online by Cambridge University Press

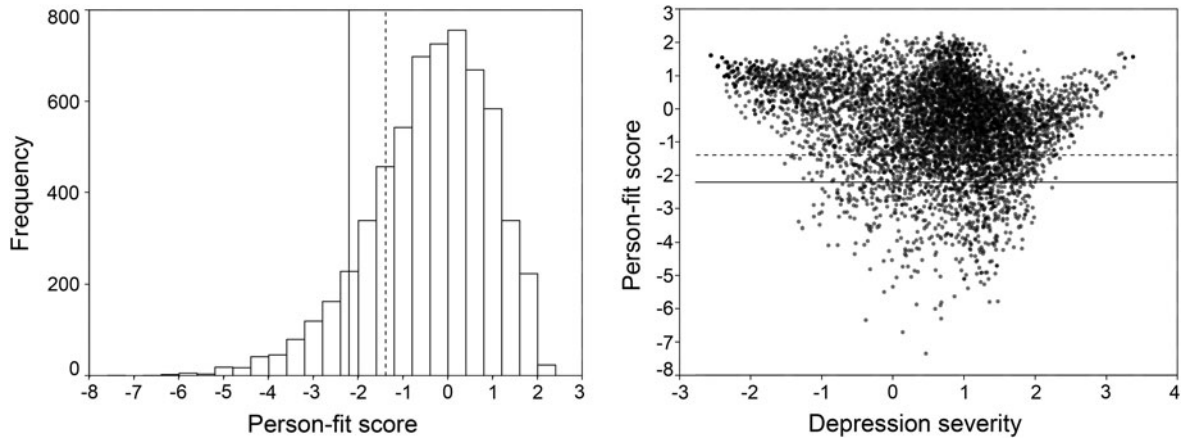

Fig. 1. Person-fit score distribution (left) and across different levels of depression severity (right) with the dotted line representing the 5% cutoff score ($l_z < -1.39$) and the solid line the 1% cutoff score ($l_z < -2.21$).

1% level. Comparably, in those with a primary mood diagnosis ($n = 512$) person-fit scores were below the 5% and 1% level, respectively, for 122 (23.8%) and 63 (12.1%) patients. For further analyses, the more conservative 1% significance level was taken as cutoff ($l_z < -2.21$), indicative of inconsistent symptom patterns.

Mean item scores of the inconsistent symptom patterns differed substantially from the patterns of typical responders (online Supplementary Fig. 3) and were characterized by lower scores on 'anxious', 'somatic complaint', 'sympathetic arousal', and 'sensitivity' and higher levels of 'reactivity of mood', 'involvement', 'enjoyment', and 'psychomotor slowing' (mean differences significant at $p < 0.001$). To investigate associations of being an inconsistent responder (yes/no) with external variables, we performed a multivariate logistic regression, with gender and primary clinical diagnoses as binary predictors, age as continuous predictor, and IDS-SR sum score as covariate adjusting for depression severity. Patients with inconsistent symptom patterns were older (mean age of 43.1 *v.* 40.4 years in typical responders; Cohen's $d = 0.17$; $z = 2.74$; $p < 0.01$), more often male (59.9% *v.* 46.2% in typical responders; OR = 1.6; $z = 3.41$; $p < 0.01$), and less often diagnosed with a primary clinical diagnosis of an anxiety disorder (9.2% *v.* 16.6% in typical responders; OR = 0.6; $z = -2.34$; $p < 0.05$). Presence of other clinical diagnoses was not a significant predictor of inconsistent response patterns.

Person fit on repeated measurements

Person-fit scores were positively correlated between first and second measurements ($r = 0.45$). Interestingly, several patients had stable inconsistent profiles over multiple measurements (online Supplementary Fig. 4). Anecdotally, one patient (#1492) with a primary diagnosis of MDD (first episode) had 24 measurements, of which 21 flagged as inconsistent (person-fit range: −1.1 to −5.1; mean $l_z = -3.12$).

To gain more insight into possible consistency of inconsistent response behavior across measurements, repeated measurements of six randomly selected patients are plotted in Fig. 2. The two patients (#1492 and #1098) with >90% inconsistent measurements showed inconsistencies that were similar across measurements, suggesting a systematic cause underlying the responses not reflective of depression. Patterns of both patients showed high scores on several severe symptoms (e.g. 'Psychomotor slowing') with many mild symptoms absent (e.g. 'Involvement'). Both examples also suggest that poor person fit is not simply caused by a single reporting of a severe symptom without reporting of milder symptoms, but instead poor person-fit scores are observed when many of these deviations are present.

Alternatively, patient #1378 showed a pattern with normal first measurements, reporting many symptoms and an IDS score of 49 at intake. In later measurements (5 months later), depressive severity improved to a score of 24, but subsequent measurements remained around this score and were flagged as inconsistent with predominantly symptoms of 'Enjoyment', 'Sexual Interest', and 'Energy' still reported. This suggests that although the depression improved overall, residual symptoms remained that led to inconsistent patterns with depression severity potentially overestimated.

Qualitative follow-up assessments on person fit

Qualitative assessments of three psychiatrists on the potential causes of inconsistent patterns for 20 of their patients are summarized in Table 2. Patients had an average age of 46.2 (range: 21–80), showed mild to severe depression severity (mean IDS-SR: 36; range: 17–52) and had poor person-fit scores (mean $l_z = -3.2$; range: −2.2 to −4.7).

Psychiatrists reported to be well acquainted with the patient in 17 of the 20 cases, and reasonably well in the remaining three cases. For 14 of the 20 patients, the inconsistent symptom pattern was conform the clinical impression. In most cases, the explanation for the inconsistent pattern was that symptoms were experienced for other reasons than MDD. For example, psychiatrists mentioned complex comorbidity (e.g. #379), somatic complaints (e.g. #1378) or the presence of isolated symptoms (e.g. #1975) as possible explanations. For six patients, the inconsistent pattern was, in retrospect, not in agreement with the psychiatrists' clinical impressions. Here, an alternative explanation could be offered in five cases. These explanations pointed at high levels of overall psychiatric distress with clinically significant problems besides depression (e.g. #2898), and motivational or concentration problems (e.g. #1723). Overall, poor person fit could be linked to a diverse range of possible underlying causes. Three illustrative cases are discussed in more detail below (patients #926, #1531, and #187).

Patient #926 had a depression with somatic comorbidity scoring disproportional high on somatic symptoms (e.g. gastrointestinal problems and low energy) not reflective of depression severity and inflating total score. In this case, information obtained from

https://doi.org/10.1017/S003329171700335X Published online by Cambridge University Press

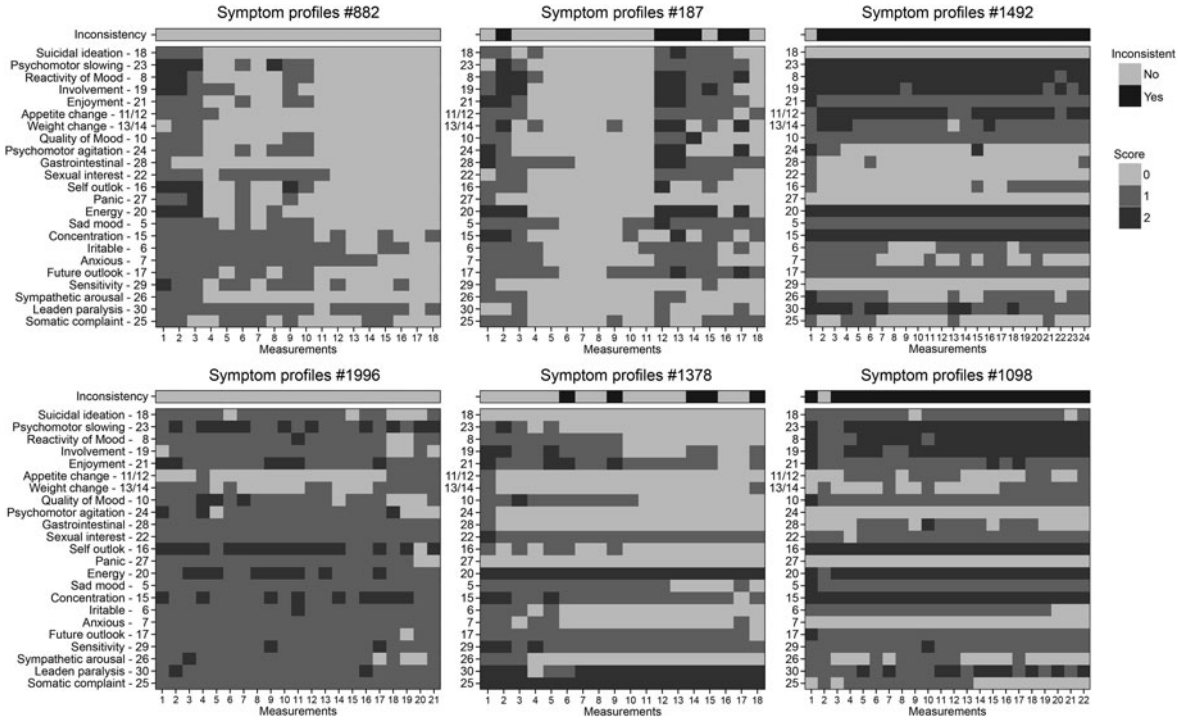

Fig. 2. Symptom profiles of six patients with repeated measurements with depressive symptoms ordered from mild (bottom) to severe (top) based on severity thresholds obtained from the IRT model. Each block represents a score (0,1,2) reported on the symptom at the time of measurement. For the first two patients on the left (#1996 and #892) no measurements are flagged as inconsistent, for the next two patients in the middle (#1378 and #187) <25% are flagged as inconsistent, and the last two patients on the right (#1492 and #1098) have >90% measurements flagged as inconsistent.

person fit may warn the psychiatrist to be careful when interpreting the total score.

Interestingly, patient #1531 had a wish to be discharged and presumably might have pretended to be better than he actually was. This could lead to an inconsistent pattern where the patient reported improvement on some obvious depressive symptoms and not on others. The two prior measurements showed extreme IDS-SR scores of 57 and 64 (indicative of severe depression in the past month; good person fit $l_z > 0.7$), strengthening the psychiatrists' interpretation of possible under-estimation of depression severity at the inconsistent measurement (IDS-SR of 17).

Patient #187 reported few symptoms with a moderate IDS-SR score of 26, but did report severe symptoms like psychomotor agitation and psychomotor slowing. The patient showed abnormalities during neurological examination, with problems in information processing and slowness of thought (bradyphrenia). For this patient the IDS-SR may not measure depression symptomatology (i.e. severity) but rather reflect neurological defects, showing the potential extra diagnostic information of person fit for a clinician.

Psychiatrists were also asked about the potential clinical usefulness of a person-fit alert if offered at the time of measurement (Table 3). For 13 of the 20 patients psychiatrists indicated that the alert would have been of direct clinical use. For the remaining seven patients, in five cases the inconsistency was already expected at the time and conform clinical impression, one patient (#187; described above) was referred to a neurologist, and one patient (#575) was reported to have come for planned specialized treatment, for which the measurement would not have led to changes in treatment policy.

Psychiatrists indicated that for 12 of the 20 patients the person-fit alert would have led to new insights: it could have increased understanding (#2816), alerted the psychiatrist to things they potentially missed (#1975) and could have led to further diagnostic examination (#2898). For the other eight patients, the inconsistency was already expected and conformed to the clinical picture. Still, the psychiatrists pointed out that the alert would have been a useful confirmation of the clinical impression that they had of the patient, helping IDS-SR score interpretation (e.g. patient #1487).

With regard to potential actions taken after a person-fit alert, psychiatrists reported for 14 of 20 patients it would be reason to inspect item scores. In addition, for 15 of 20 patients the alert would have been a reason to discuss possible inconsistencies with the patient and could serve as a useful starting point for a discussion with a patient on specific diagnostic issues (e.g. the possible downplaying of depression for patient #2816).

Discussion

This study aimed to investigate the clinical meaning and usefulness of inconsistent symptom profiles on IDS-SR depression severity assessments in a naturalistic clinical setting. Depressive symptoms reported by patients who completed the IDS-SR in a specialized care setting were investigated on inconsistencies by means of a data-driven approach based on person-fit statistics. Depressive symptom patterns identified as inconsistent were analyzed on both intake and repeated measurements. These results were qualitatively followed up among three psychiatrists on 20 of their randomly selected patients with inconsistent profiles, to get more insight into potential underlying causes and to evaluate the clinical use of person-fit statistics for psychiatrists.

https://doi.org/10.1017/S003329171700335X Published online by Cambridge University Press

Table 2. Explanation of psychiatrists for the inconsistency of reported symptom patterns of 20 of their patients

ID	Age	l_z	IDS	Familiar with patient	Conform clinical impression[a]	Severity estimation[b]	Explanation psychiatrist on potential cause
187	63	−4.3	26	Reasonable	Yes	–	Problems in information processing, bradyphrenia, abnormalities in neurological examination and referred to neurologist
379	43	−3.2	41	Very good	Yes	Under	Depression with complex comorbidity causing very high level of suffering at time of measurement
575	64	−3.8	39	Reasonable	–	Good	Presence of isolated symptoms
766	37	−3.2	48	Very good	Yes	Good	Severe depression with complex comorbidity including anxiety and dissociation
837	51	−3.5	21	Good	No	Good	No explanation
926	53	−4.2	43	Good	Yes	Over	Somatic comorbidity, many physical complaints especially pain
1339	46	−3.5	41	Very good	Yes	Under	Motivation/concentration problems in an episode of severe decompensation
1378	51	−2.3	24	Good	Yes	Over	Many somatic complaints leading to a higher score than expected based on patient's mood
1487	51	−2.8	52	Good	Yes	Over	High psychiatric distress with comorbid anxiety and catastrophic interpretation of pain symptoms, causing patient to be desperate about the future and suicidal
1531	71	−2.2	17	Good	No	Under	Patient wanted to be discharged and might have pretended to be better, although patient showed some clinical improvement in the week before
1543	49	−2.4	28	Very good	Yes	Under	Patient inexplicably improved during a wash-out phase (no medication), which was surprising as the patient still seemed mentally unstable and quite ill
1704	52	−3.2	48	Good	Yes	Under	Patient has bipolar depression with more psychomotor retardation and energy loss. Possibly, [the patient] was more depressed than reflected by the questionnaire as [the patient] showed limited illness awareness
1723	27	−2.4	28	Good	No	Under	Motivation/concentration problems
1975	49	−3.2	50	Very good	Yes	Good	Presence of isolated symptoms and high psychiatric distress
2541	20	−3.8	28	Good	Yes	Over	Motivation/concentration problems
2775	48	−3.9	32	Reasonable	Yes	Good	Comorbid autism spectrum disorder (PDD-NOS)
2816	80	−4.7	35	Very good	Yes	Under	Patient was treated with eskatamine in a terminal phase and later deceased through euthanasia as a result of total despair. Patient showed a tendency to trivialize his depression
2898	57	−2.5	38	Very good	No	Under	High psychiatric distress with comorbid anxiety disorder
3049	52	−2.5	47	Very good	Yes	Over	Exaggerates or feigns symptoms
3058	47	−3.3	39	Very good	No	Good	High psychiatric distress

[a]Psychiatrists were asked 'Does the inconsistency alert correspond with your own clinical impression of the patient?'.
[b]Psychiatrists were asked 'Was the total score an under-, over-, or good estimation of the severity of depression?'.

Poor person fit was frequently observed, with 12.8% of patients identified with inconsistent symptom patterns using a conservative 1% significance level (26.6% at 5% level). This is higher than previous studies reporting rates of 6.8–14% in clinical samples (Woods *et al.* 2008; Conijn *et al.* 2015; Wanders *et al.* 2015*a*; Wardenaar *et al.* 2015), but in line with the expectation that patients in specialized care present with diverse and complex psychopathology (Groenewold *et al.* 2013) and experience depressive symptoms for other reasons than depression (Wanders *et al.* 2015*b*). Furthermore, the IDS-SR was used to screen for the presence of MDD: only 25.1% of patients actually had a primary clinical diagnosis of MDD at the time of assessment, adding to the heterogeneity of the sample. As a result, it is not surprising that many patients reported symptom patterns that deviate from the typical structure of depressive symptoms, making person fit a potentially valuable source of information in this clinical setting.

https://doi.org/10.1017/S003329171700335X Published online by Cambridge University Press

Table 3. Clinical usefulness of the person-fit alert according to psychiatrists regarding 20 of their patients with inconsistent symptom patterns

ID	Age	l_z	IDS	Useful alert[a]	New insights[b]	Inspect item scores	Discuss with patient	Explanation psychiatrist on clinical usefulness
187	63	−4.3	26	No	No	No	No	No explanation given [patient was referred to neurologist]
379	43	−3.2	41	Yes	Yes	Yes	Yes	Clinically useful, if it becomes clear how this was manifested in the response behavior
575	64	−3.8	39	No	Yes	Yes	Yes	Patient came only for specific chronotherapy and an alert of inconsistency would not have led to changes in [the used treatment] policy. Would have led to a further discussion with patient
766	37	−3.2	48	Yes	Yes	Yes	Yes	Could be helpful, if it would provide more clarity about the nature of the inconsistency
837	51	−3.5	21	Yes	Yes	Yes	Yes	Possibly insightful, [psychiatrist says] to be curious on which domains the inconsistency occurred
926	53	−4.2	43	No	No	Yes	No	Inconsistency was already expected
1339	46	−3.5	41	Yes	No	Yes	Yes	Would have led to further discussion with patient. [Psychiatrist says] the inconsistency fits within the clinical picture of severe problems
1378	51	−2.3	24	No	No	No	No	The inconsistent response behavior was expected, as to us the patient had shown good clinical improvement in mood
1487	51	−2.8	52	Yes	Yes	Yes	Yes	Depending on where the inconsistency is found; if the patient is very suicidal and anxious but has lower depression severity, [the report] would fit with [the psychiatrist's] impression and could help to interpret the high IDS score, which [the psychiatrist] finds clinically incorrect
1531	71	−2.2	17	Yes	Yes	Yes	Yes	Depending on why the measurement was inconsistent it could lead to further discussion with the patient especially given his strong wish to be discharged
1543	49	−2.4	28	Yes	Yes	Yes	Yes	Could have provided more insight. Patient later deteriorated and this could perhaps been detected as a result of the person-fit alert
1704	52	−3.2	48	No	No	No	No	Patient was clinically clearly ill
1723	27	−2.4	28	No	No	No	No	It was clear that this measurement was aberrant since other measures were considerable higher, and this corresponded better with our observations during treatment
1975	49	−3.2	50	Yes	Yes	Yes	Yes	[Psychiatrist says] it would make him alert and is reason for further discussion. Possibly, [psychiatrist] missed something
2541	20	−3.8	28	Yes	Yes	Yes	Yes	Possibly insightful
2775	48	−3.9	32	No	No	No	No	Fits within the clinical picture of comorbidity
2816	80	−4.7	35	Yes	Yes	No	Yes	Help to increase understanding, and could have been a reason to discuss despair and the downplaying of his depression
2898	57	−2.5	38	Yes	Yes	Yes	Yes	Could lead to further diagnostic examination
3049	52	−2.5	47	Yes	Yes	Yes	Yes	Would strengthen the current clinical picture
3058	47	−3.3	39	Yes	No	Yes	Yes	Would confirm that the severity indeed is high

[a]Psychiatrists were asked: 'Would the alert of possible inconsistency be useful in this case for you as a psychiatrist?'.
[b]Psychiatrists were asked: 'Could the alert of possible inconsistency have led to new insights?'.

In contrast to other applications of person-fit research (e.g. educational assessment and selection), in clinical research the causes underlying poor person fit are often of systematic nature. These causes may include the presence of comorbidity like anxiety or somatic complaints (Wanders *et al.* 2015*a*; Wardenaar *et al.* 2015) and cognitive difficulties (Conijn *et al.* 2017), but also entail the presence of a different primary disorder, such as a neurological disorder (e.g. patient #187). Interestingly,

https://doi.org/10.1017/S003329171700335X Published online by Cambridge University Press

the systematic influence of these factors on patients' response patterns was also supported by the observation that inconsistent response behavior was rather stable across measurements (e.g. patient #1492). One way to better identify different causes of inconsistent symptom reports would be to utilize and develop more objective measures of the factors found to be associated with the quality of the responses. For example, studies associated self-report quality indicators like an extreme or agreement response style (van Herk *et al.* 2004; Conijn *et al.* 2016), interviewer evaluation of language problems (Conijn *et al.* 2016), and external indicators of cognitive difficulties (Conijn *et al.* 2017) with poorer person fit. Another promising approach would be to analyze the patterns of time spent by a patient on answering each item in computer-based assessments (Marianti *et al.* 2014). Careless and unmotivated behavior would be expected to cause aberrant response time patterns (Marianti *et al.* 2014), providing an opportunity to obtain more objective information on person misfit.

In the current study, we distinguished between several clinical applications of person fit based on the results of the qualitative psychiatrist assessments. First, an alert of inconsistency can serve as a warning signal for the psychiatrist that the total score is not a good representation of depression severity and should be interpreted with caution. In some cases, psychiatrists indicated to be curious on what symptoms caused the inconsistencies. Inspection of item scores combined with available clinical information, following a discussion with the patient could clarify such discrepancies. Alternatively, clinicians could be provided with additional information together with the person-fit alert. The system could adaptively respond to inconsistencies by administering additional items or questionnaires (Liu & Yu, 2011), or could automatically perform follow-up analyses on differences between expected and observed item scores (Ferrando, 2015) to detect deviation trends across symptoms and identify possible sources of symptom-level misfit.

Second, a person-fit alert could confirm the current clinical impression that the psychiatrist has of a patient, especially in cases where a typical profile is not expected *a priori*. For example, in the case of patient #1378, where the psychiatrist saw a patient clinically improving, but this improvement was not reflected in the IDS-SR total scores. Retrospectively the measurement was identified as inconsistent, which could have served as a confirmation for the psychiatrist's impression and could have supported the clinical decisions made.

Third, an alert of a purely statistical method could serve as an opportunity to discuss particular issues with a patient. For example, in case of patient #2816, where the suspicion was that the patient was downplaying his depression, the person-fit alert could have served as a starting point to discuss a topic that might be difficult to talk about without directly blaming the patient of downplaying. In future research, the improvement of care when this type of feedback is provided should be studied in a pragmatic cluster randomized trial, providing person-fit feedback to one experimental group of clinicians and maintaining care as usual in the other experimental group. In such trials, several outcomes could be compared, including additional diagnostic work-up, patient and clinician satisfaction, and the quality of the working-alliance. Ultimately, pragmatic trials could be used to investigate the effects of providing automated person-fit feedback on treatment outcome and cost-effectiveness of care.

This study had several limitations. Although this is the first study in which person-fit statistics were implemented and interpreted in the context of a real clinical setting, the design was still retrospective. Therefore, the current results should be seen as a proof-of-principle. Further prospective studies are needed, where person fit is implemented in real time and feedback to clinicians is given on-the-fly at the moment of assessment. An additional limitation is that the psychiatrists knew that the person-fit scores of their patients were low before they gave their feedback (they were not blinded or provided with sham-cases of poor person fit). Also, it is possible that the three psychiatrists agreed to participate in the current study because they had affinity with research and felt more positive toward the use of statistics/technology in depression assessment than other psychiatrists. This may have resulted in a positive bias toward the clinical use of these statistics. In addition, we calibrated the group-based model on a well-defined but external sample (NESDA; Penninx *et al.* 2008). An alternative would have been to calibrate the model on only patients with a diagnosis of MDD or to obtain a more homogenous subsample with the use of mixture modeling (Rupp, 2013).

The promising results of recent person-fit studies in clinical assessment (e.g. Woods *et al.* 2008; Conrad *et al.* 2010; Conijn *et al.* 2015; Wanders *et al.* 2015*a*; Wardenaar *et al.* 2015) raised important questions on the causes behind inconsistent patterns and whether implementation of person-fit statistics could be of clinical use. The current study affirmed that there are real clinical causes behind inconsistent symptom profiles that give poor person-fit scores a clinical interpretation. Above all, the feedback collected among psychiatrists suggested that person-fit alerts could be highly informative for clinicians when interpreting depression assessments, and of valuable support in clinical decision-making. In this context, all relevant information is summarized to guide clinical decisions (Puschner *et al.* 2010) and a person-fit message should be seen as a piece of extra information on top of the regularly used severity sum scores. With evidence converging on the usefulness of person-fit statistics, routine assessments taking place with automated systems (Lambert & Shimokawa, 2011), person-fit software (e.g. Ferrando & Lorenzo, 2000; Tendeiro *et al.* 2016) and nontechnical tutorials being available (e.g. Meijer *et al.* 2016), person fit is ready for on-the-fly implementation in depression assessment.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S003329171700335X

Acknowledgements. The current study was supported by a VICI grant (no: 91812607) received by Peter de Jonge from the Netherlands organization for Scientific research (NWO-ZonMW).

Declaration of Interest. None.

References

Conijn JM, Emons WH, De Jong K and Sijtsma K (2015). Detecting and explaining aberrant responding to the outcome questionnaire-45. *Assessment* **22**, 513–524.

Conijn JM, Spinhoven P, Meijer RR and Lamers F (2016). Person misfit on the inventory of depressive symptomatology: low quality self-report or true atypical symptom profile? *International Journal of Methods in Psychiatric Research*. doi: 10.1002/mpr.1548.

Conijn JM, van der Ark LA and Spinhoven P (2017). Satisficing in mental health care patients: the effect of cognitive symptoms on self-report data quality. *Assessment*. doi: 1073191117714557.

Conrad KJ, Bezruczko N, Chan YF, Riley B, Diamond G and Dennis ML (2010). Screening for atypical suicide risk with person fit statistics among

people presenting to alcohol and other drug treatment. *Drug and Alcohol Dependence* **106**, 92–100.

Drasgow F, Levine MV and Williams EA (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology* **38**, 67–86.

Embretson SE and Reise SP (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum: Mahwah, NJ.

Embretson SE and Reise SP (in press). *Item Response Theory*. Routledge Taylor & Francis: New York, NY.

Ferrando PJ (2015). Assessing person fit in typical-response measures. In *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (ed. S.P. Reise and D. A. Revicki), pp. 128–155. Routledge: New York, NY.

Ferrando PJ and Lorenzo U (2000). WPerfit: a program for computing parametric person-fit statistics and plotting person response curves. *Educational and Psychological Measurement* **60**, 479–487.

Groenewold NA, Doornbos B, Zuidersma M, *et al.* (2013). Comparing cognitive and somatic symptoms of depression in myocardial infarction patients and depressed patients in primary and mental health care. *Plos One* **8**, e53859.

Lambert MJ and Shimokawa K (2011). Collecting client feedback. *Psychotherapy* **48**, 72–79.

Liu MT and Yu PT (2011). Aberrant learning achievement detection based on person-fit statistics in personalized e-learning systems. *Journal of Educational Technology and Society* **14**, 107–120.

Marianti S, Fox JP, Avetisyan M, Veldkamp BP and Tijmstra J (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics* **39**, 426–451.

Meijer RR (2003). Diagnostic item score patterns on a test using IRT based person-fit statistics. *Psychological Methods* **8**, 72–87.

Meijer RR, Egberink IJL, Emons WHM and Sijtsma K (2008). Detection and validation of unscalable item score patterns using item response theory: an illustration with Harter's Self-Perception Profile for Children. *Journal of Personality Assessment* **90**, 227–238.

Meijer RR, Niessen ASM and Tendeiro JN (2016). 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**, 52–62.

Meijer RR and Sijtsma K (2001). Methodology review: evaluating person fit. *Applied Psychological Measurement* **25**, 107–135.

Pallant JF and Tennant A (2007). An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology* **46**, 1–18.

Penninx BWJH, Beekman ATF, Smit JH, Zitman FG, Nolen WA, Spinhoven P, Cuijpers P, De Jong PJ, Van Marwijk HWJ, Assendelft WJJ, Van Der Meer K, Verhaak P, Wensing M, De Graaf R, Hoogendijk WJ, Ormel J and Van Dyck R (2008). The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *International Journal of Methods in Psychiatric Research* **17**, 121–140.

Puschner B, Steffen S, Slade M, Kaliniecka H, Maj M, Fiorillo A, Munk-Jørgensen P, Larsen JI, Égerházi A, Nemes Z, Rössler W, Kawohl W and Becker T (2010). Clinical decision making and outcome in routine care for people with severe mental illness (CEDAR): study protocol. *BMC Psychiatry* **10**, 90.

Reise SP and Waller NG (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology* **5**, 27–48.

Rizopoulos D (2006). Ltm: an R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software* **17**, 1–25.

Rupp AA (2013). A systematic review of the methodology for person fit research in item response theory: lessons about generalizability of inferences from the design of simulation studies. *Psychological Test Assessment Modeling* **55**, 3–38.

Rush AJ, Gullion CM, Basco MR, Jarrett RB and Trivedi MH (1996). The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychological Medicine* **26**, 477–486.

Samejima F (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement* **34**, 100.

Tendeiro JN, Meijer RR and Niessen ASM (2016). Perfit: an R package for person-fit analysis in IRT. *Journal of Statistical Software* **74**, 1–27.

Thomas ML (2011). The value of item response theory in clinical assessment: a review. *Assessment* **18**, 291–307.

Van Herk H, Poortinga YH and Verhallen TM (2004). Response styles in rating scales: evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology* **35**, 346–360.

Wanders RBK, Wardenaar KJ, Kessler RC, Penninx BWJH, Meijer RR and De Jonge P (2015*b*). Differential reporting of depressive symptoms across distinct clinical subpopulations: what difference does it make? *Journal of Psychosomatic Research* **78**, 130–136.

Wanders RBK, Wardenaar KJ, Penninx BWJH, Meijer RR and De Jonge P (2015*a*). Data-driven atypical profiles of depressive symptoms: identification and validation in a large cohort. *Journal of Affective Disorders* **180**, 36–43.

Wardenaar KJ, Wanders RBK, Roest AM, Meijer RR and 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**, 130–142.

Woods CM, Oltmanns TF and 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**, 159–168.

https://doi.org/10.1017/S003329171700335X Published online by Cambridge University Press