OXFORD

Research Methods

Evaluating the validity of the French version of the Four-Dimensional Symptom Questionnaire with differential item functioning analysis

Juliette Chambe[a],*, Jean-Yves Le Reste[b], Hubert Maisonneuve[c], Anne-Elisabeth Sanselme[a], Julienne Oho-Mpondo[a], Patrice Nabbe[b] and Berend Terluin[d]

[a]Department of General Practice, University of Strasbourg, Strasbourg, France, [b]ERCR SPURBO, Department of General Practice, Université de Bretagne Occidentale, Brest, France, [c]Primary Care Unit, Faculty of Medicine, University of Geneva, Geneva, Switzerland and [d]Department of General Practice and Elderly Care Medicine, EMGO Institute for Health and Care Research, VU University Medical Centre, Amsterdam, The Netherlands.

*Correspondence to Juliette Chambe, Department of General Practice, University of Strasbourg, 4 rue Kirschleger, Strasbourg 67000, France; E-mail: juliette.chambe@unistra.fr

Abstract

Background. GPs are daily confronted with mental disorders and psychosomatic problems. The Four-Dimensional Symptom Questionnaire (4DSQ), measuring distress, depression, anxiety and somatization, was purposively developed for primary care. It has been translated into 12 languages and is commonly used in several countries. It was translated into French in 2008, by forward and backward translation, but it has not been validated for a primary care population.
Aim. This study aimed to establish whether the French 4DSQ measured the same constructs in the same way as the original Dutch 4DSQ.
Method. Two samples of French general practice patients were recruited during routine care to obtain as much variability as possible. One sample included consecutive patients, from the waiting room of rural GPs, over a period of 2 weeks and the other sample included patients with suspected psychological problems or unexplained symptoms. This population was compared to a matched Dutch sample using confirmatory factor analysis (CFA) and differential item functioning (DIF) analysis.
Results. A total of 231 patients, from 15 French GPs, completed the questionnaire (Dutch reference group: 231). Mean age was 42.9 years (Dutch: 42.1); females numbered 71% in both samples. The multigroup CFA assessed configural invariance of one-factor models per 4DSQ scale. Thirteen of the total of 50 items in the 4DSQ, in three scales, were detected with DIF. However, DIF did not impact on the scale scores.
Conclusion. French 4DSQ scales have the same latent structures and measure the same traits as the original Dutch 4DSQ.

Key words: Medically unexplained symptoms, mental health, primary care, screening.

Introduction

Self-administered questionnaires are widely used to evaluate mental disorders in hospital. For example, the Beck Depression Inventory and the Beck Anxiety Inventory are commonly used either for following up patients or for research. They are translated and validated in many languages. Unfortunately, these questionnaires received

little development or validation within primary care. These questionnaires are actually of limited interest in primary care, as they were designed to screen and follow up severe psychiatric disease. Few questionnaires were specifically developed for general practice, such as the Mini International Neuropsychiatric Interview test for depression (1). In general practice, around a third of the consultations are related to psychological problems, including both 'classic' psychiatric disorders, subsyndromal and psychosocial problems (2). These problems are at the same time very common, and difficult for the GP to investigate, as they are often mixed with other symptoms, and difficult for the patient to admit to. Use of questionnaires can help to detect, evaluate or follow these patients, especially to make the distinction between psychiatric disorders and psychosocial problems. Self-administered questionnaires can be more efficient to use in general practice and, in addition, could help the patient to consider the psychosocial source of their physical disorders (3). For those reasons, Terluin, a Dutch primary care doctor and researcher, developed a questionnaire, the Four-Dimensional Symptom Questionnaire (4DSQ), which measures four dimensions of psychological problems commonly encountered in a GP's practice: distress, anxiety, depression and somatization (4). It helps patient and physician to acknowledge psychological distress when they are confronted with unexplained symptoms. This 50-item questionnaire has been translated into 12 languages and has already had Dutch, Polish and English versions validated (5,6). It was translated into French using a forward and backward method. However, it is known that intended meaning could be lost or influenced by cultural variations in the expression of attitudes and emotions. This could give rise to an unintended change of meaning within the translated questionnaire. We proceeded to a validation process of the French 4DSQ, to ensure that it worked in the same way as the original Dutch questionnaire.

Research question

Does the French 4DSQ measure the same constructs in the same way as the original Dutch 4DSQ questionnaire? To answer this question, we took the approach of assessing measurement equivalence, i.e. the equivalence of the measurement models of the Dutch and French 4DSQ versions. When the French 4DSQ demonstrates equivalence with the Dutch 4DSQ, it can be assumed that the validity and reliability of the Dutch 4DSQ will also apply to the French 4DSQ.

Method

Design and participants

Data were obtained from two samples of French General Practice patients, drawn from two different parts of France. Both groups consisted of adult patients (aged 18–64 years), attending 12 GPs in 2013. The patients were recruited during routine care. Patients were selected using two different methods in order to obtain a wide variability of responses. In one sample (Group 1), questionnaires were given out to consecutive patients by five GPs from rural areas, in their waiting room, over a 2-week period. In the other sample (Group 2), seven GPs were asked to recruit, preferably, patients with (suspected) psychological problems, such as distress, depression or anxiety, or medically unexplained symptoms, defined as persistent symptoms for which no medical causes were found, such as pain, dizziness and gastrointestinal troubles. Each of the GPs in this sample, from urban and rural areas, was asked to issue 10 questionnaires. These two ways of selecting were chosen on purpose to obtain as wide a variability as possible. For practical reasons, a nil response could not be recorded. The patients were fully informed about the study's aim.

Participation was voluntary and patients consented by returning the completed questionnaire. Data were immediately anonymized.

For comparison with the French data, 4DSQ data from Dutch patients, matched on gender and age (10-year groups), were drawn from a large sample of 1114 patients with (suspected) mental health problems or medically unexplained symptoms, visiting their GP in a primary care health centre in Almere, in the Netherlands. The Dutch 4DSQ data had been collected during routine care.

The questionnaire

The 4DSQ is a 50-item self-report questionnaire. Those items explore four scales: distress, depression, anxiety and somatization. First, depression (6 items) and anxiety (12 items) were evaluated from specific symptoms of depressive and anxiety disorders (4 items). The 16 items on the distress scale measure the global response to people's stress. The scale evaluates various aspects: associated with work, family, life events or psychosocial difficulties of any kind (3 items). Somatization is measured by symptoms associated with somatic distress (16 items). Response choices follow a Likert scale: 0 for 'no', 1 for 'sometimes' and 2 for the other categories: 'regularly', 'often', 'very often or constantly'. The item scores are totalled within each dimension scale. The interpretation of each score is worked out according to validated cut-off points: normal, moderate and severe scores. Finally, these scales inform the professional's decision making. The 4DSQ is freely available for non-commercial use at: www. emgo.nl/researchtools/4dsq.asp.

The first French translation was created in 2008. Two GPs, French native speakers, drafted the first French translation, which was then back-translated by an independent Dutch native speaker and professional translator. Discrepancies between the back translation and the original Dutch text were resolved by discussion with both translators.

Missing item scores were imputed using the response function method, provided that no more than half of the item scores of a scale were missing (7).

Analyses

First, essential uni-dimensionality was evaluated using multigroup confirmatory factor analysis (CFA) as implemented in the R package 'lavaan' (8). We fitted, for each of the 4DSQ scales, one-factor models allowing residual correlations between item pairs sharing specific content. The item scores were treated as ordered indicators. Fit measures indicating adequate fit included a confirmatory fit index >0.95, Tucker–Lewis index >0.95 and root mean square error of approximation <0.06 (9).

Second, to assess measurement equivalence, we used differential item functioning (DIF) analysis (10). DIF analysis examines whether an item demonstrates the same 'function' across different groups (e.g. French and Dutch patients), as evidenced by the item's relationship to the latent trait (e.g. depression) the item is supposed to be an indicator of. The item–trait relationship is expressed in the correlation between item and trait and in the threshold for being endorsed. In the absence of DIF, the French items and their original Dutch counterpart items possess the same correlations with the underlying latent trait and the same thresholds for endorsement. DIF is present when different groups (e.g. French and Dutch patients) have different probabilities of endorsing the response options of an item, while having the same position on the underlying trait that the item is supposed to measure (11). Following the advice of Hambleton (12), we used two different methods: the non-parametric Mantel–Haenszel (M–H) method (13), as implemented in the statistical program

jMetrik (14), and the parametric ordinal logistic regression (OLR) method, as implemented in the R package 'lordif' (15). DIF was considered present when, in the M–H analysis, the standardized mean difference was >0.1 and $P < 0.01$ or, in the OLR analysis, the R^2 difference was >0.02 and $P < 0.01$.

Third, we assessed the impact of DIF on the scale scores. This is of greater importance because in research and clinical practice the scores are usually interpreted on the scale score level and not on the individual item level. The impact of DIF was evaluated by comparing the raw scale scores with item response theory (IRT) scale scores produced by Rasch analysis, using the DIF-free items as anchors to fit all items and patients on the same scales. The IRT score was thus adjusted for the effect of DIF in certain items and reflected the true (e.g. DIF-free) trait score. Rasch analysis was performed using jMetrik (14).

Results

The French group consisted of 179 primary care patients from Group 1 and 52 primary care patients from Group 2 (total $n = 231$). The percentage of females was 71%. The mean age was 42.9 years (SD = 11.7). The Dutch reference group consisted of 231 primary care patients, 71% of whom were female, and the mean age was 42.1 years (SD = 11.6). In the French sample, 15% had one or more missing item scores; in the Dutch sample, this was the case in 23% of the patients. The percentage of missing item scores was very low: 0.5% in the French sample and 0.7% in the Dutch sample. The missing scores were successfully imputed.

The results of the multigroup CFA are presented in Table 1. All 4DSQ scales demonstrated adequate fit of the one-factor model. Residual correlations, to improve the model fit, were allowed for two item pairs of the distress scale and two item pairs and one item triplet of the somatization scale. These items arguably shared common content (see the footnote to Table 1).

Table 2 presents the results of the DIF analyses. The distress scale turned out to be free of DIF. A total of 13 items were flagged for DIF by either method. The M–H method detected seven DIF items and the OLR method detected nine items. Only three DIF items were detected by both methods. Seven items were found to be less severe for French patients. That means that French patients had a lower threshold than Dutch patients for endorsing these items. As an illustration, Figure 1 (left panel) shows the 'item response function' of Item 3, an item of the somatization scale. French patients endorsed this item at a lower level of somatization than Dutch patients. French patients needed less severe somatization to endorse this item than

Dutch patients. Apparently, 'des malaises' was a less severe symptom of somatization than 'flauw vallen' (in English: 'fainting'). On the other hand, six items appeared to be more severe for French patients than for Dutch patients. This means that the item represented a more severe symptom. Figure 1 (right panel) shows an item that was more severe for French patients. Apparently, 'comme ça, un saisissement soudain' represented a more severe level of anxiety than 'zomaar plotseling schrikken' (in English: 'sudden fright for no reason'). An item that is less severe might lead to higher scores than can be explained by the true level of somatization or anxiety, whereas an item that is more severe can lead to a lower score.

The Rasch analyses revealed that the DIF-laden French items generally fitted the Rasch scale model equally well or even better than the Dutch counterpart items (data not shown). Figure 2 presents the graphical comparison of the raw scale scores with the DIF-free IRT score. The vertical distance between the group-specific curves proves the impact of DIF. Figure 2 demonstrates that the impact of DIF was limited to (very) high levels of depression and anxiety. In terms of the interpretation of the scores in relation to cut-off scores, the impact of DIF was negligible. It is probable that the DIF in certain opposing items cancelled each other out within the scale.

Discussion

Our results demonstrate that the French 4DSQ scales have the same latent structures and measure the same traits as the original Dutch 4DSQ. Therefore, the French 4DSQ scores convey the same meaning as the Dutch scores and the scales can use the same set of cut-off points. In this way, the same clinical decisions are reached.

Limitations

Population selection was done in two different ways: the majority of the patients ($n = 179$) were recruited consecutively in order to obtain a representative group from GPs' consulting patients. The other group of patients was recruited with a specific purpose, in order to be representative of the patients consulting for one of the four dimensions explored by the questionnaire.

Patients' sociodemographic data, other than gender and age, were not available, and we cannot check their representativeness. However, the chosen statistical analysis method does not require those data for the questionnaire validation.

Incomplete questionnaires were also analysed, using imputation; however, the completion rate was very good (99.5%).

The DIF encountered in 13 items did not have a substantial impact on the 'test functioning'. Only three of these items, all related to anxiety, were identified both by M–H and OLR methods.

French patients tended to underscore their level of anxiety especially when the (true) level of anxiety was high. This applies also to the way French patients score their level of depression. At the same time French patients tended to overscore their level of somatization, but only to a very small degree.

It is interesting to note that in English, Polish, as well as French translations, there is a DIF for the same anxiety item no. 27. 'Angoissé' (F) and 'Wystraszony' (PL) were apparently less severe than 'angstig' (NL), but 'frightened' (EN) appeared to be more severe (5,6). Some assumptions can be made to explain those differences: even when translations are close to the original meaning they can remain linguistically imperfect, especially when one word covers a wide range of meanings and strong subjective interpretation. We can also hypothesize that there are some cultural differences in meaning and that an exact equivalent is not always possible. In French, this

Table 1. Results of the multigroup CFA, assessing configural invariance of one-factor models per 4DSQ scale

4DSQ scale	CFI	TLI	RMSEA	95% CI
Distress[a]	0.997	0.996	0.046	0.035–0.057
Depression	1.000	0.999	0.042	0.000–0.077
Anxiety	1.000	1.002	0.000	0.000–0.019
Somatization[b]	0.990	0.988	0.046	0.035–0.057

CFI, comparative fit index; CI, confidence interval; RMSEA, root mean square error of approximation; TLI, Tucker–Lewis index.

[a]Correlated residual variance item pairs 47–48 (upsetting events) and 20–39 (sleep).

[b]Correlated residual variance of item pairs 15–16 (thoracic symptoms), 12–13 (gastrointestinal symptoms) and item triplet 2–4–5 (musculoskeletal symptoms).

Table 2. Items detected with DIF

Scale	Item number	Short description (French/*English*)	M–H method[a]	OLR method[b]	Direction[c]
Depression	33	mieux que vous soyez mort/*better off if you were dead*	–	0.025	+
	34	rien ne pouvait vous réjouir/*you can't enjoy anything anymore*	–	0.048	–
Anxiety	18	saisissement soudain/*sudden fright for no reason*	–0.10	0.055	–
	21	sensation indéfinissable de peur/*a vague feeling of fear*	–	0.027	–
	24	soudaine panique ou forte anxiété/*anxiety or panic attacks*	+0.15	0.023	+
	27	Angoissé/*Frightened*	+0.18	0.021	+
	45	menacé(e) par un danger inconnu/*being threatened by unknown danger*	–	0.030	–
	49	dû éviter certains endroits/*have to avoid certain places*	–	0.039	–
Somatization	1	des vertiges ou la tête qui tourne/*dizziness or feeling light-headed*	–0.12	–	–
	2	des douleurs musculaires/*painful muscles*	+0.25	–	+
	3	des malaises/*fainting*	–	0.063	+
	4	une douleur dans le cou/*neck pain*	+0.19	–	+
	9	une sensation de ballonnement/*a bloated feeling in the abdomen*	+0.21	–	+

[a]M–H method: standardized mean differences.

[b]OLR method: R^2 differences.

[c]Direction of DIF: + the item is less severe for French patients; – the item is more severe for French patients.
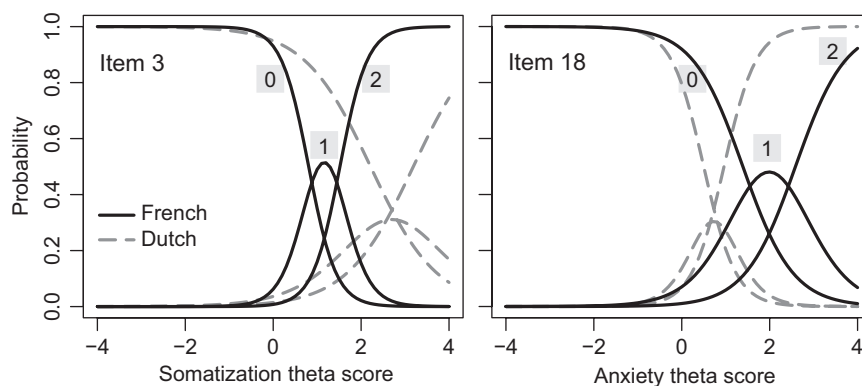
Figure 1. Item response function curves of two items demonstrating DIF comparing French and Dutch patients. The graphs display the probability of endorsing the item response options 'no' (coded '0'), 'sometimes' (coded '1') and 'regularly', 'often' or 'very often or constantly' (coded '2') as a function of the DIF-adjusted IRT theta score, which represents an approximation of the 'true' somatization or anxiety score. Item 3 ('fainting'; left panel), belonging to the somatization scale, was less severe for French patients than for Dutch patients: French patients endorsed response options '1' and '2' at less severe theta levels than Dutch patients did. Item 18 ('sudden fright for no reason'; right panel), part of the anxiety scale, was more severe for French patients than for Dutch patients: French patients endorsed response options '1' and '2' at more severe theta levels than Dutch patients did.

term has a broad and emphatic meaning, covering both body concerns and psychological feelings. This kind of variability is explored in a European study about depression and multi-morbidity (16).

Despite the fact that some translated items turned out to function differently from the original items, the scale scores of the translated French 4DSQ function in largely the same way as in the original 4DSQ.

Use of the 4DSQ in French practice

In France, no questionnaire has yet been validated for depression and anxiety. As it is the second most frequent chronic disease seen in primary care (16), it is important to have a first draft for consideration on the 4DSQ. Indeed, this questionnaire is widely and successfully used in the Netherlands.

Other international depression scales have been developed and validated for primary care (17), but those questionnaires are only able to sort anxiety and depression and do not take into account the other dimensions explored in the 4DSQ. And yet, distress and somatization feature in GPs' consultations to a huge extent: psychological problems and medically unexplained symptoms represent >20% of consultations (2).

The 4DSQ has the advantage of enabling the GPs to explore those issues, which French GPs find particularly difficult to approach, and which patients may avoid or deny. A validated tool can improve the practice in several ways: it could be used to help the physician in the diagnostic process and in the shared decision making. Then, it could be a new medium for improving patient–physician communication in those sensitive situations. This kind of tool could help to implement the biopsychosocial model in the practice (18).

Oho-Mpondo Master's thesis not only showed a good congruence of the depression scale between patients and GPs but also suggested a discrepancy in the evaluation of distress and somatization between patients and GPs (19). Those results are consistent with Schumacher's study, which showed a high level of disagreement between patients and physicians about medically unexplained symptoms (20). This questionnaire could be especially helpful at a denial stage of a psychological or psychosomatic problem. Further feasibility studies could answer this hypothesis.

The primary care system is not organized in the same way in the Netherlands as it is in France. In the Netherlands, physicians work in a team with an assistant/nurse, whereas they mostly work
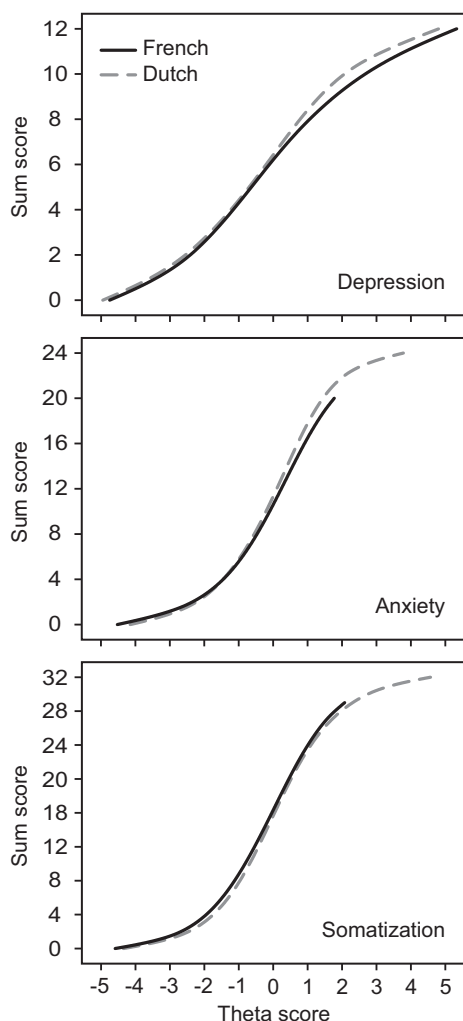
Figure 2. Impact of DIF on the raw scale scores for depression (upper panel), anxiety (middle panel) and somatization (lower panel) comparing French and Dutch patients. The graphs display the raw scale score as a function of the DIF-adjusted Rasch theta score, which represents an approximation of the 'true' depression, anxiety or somatization score. The vertical distance between the group-specific curves represents the scale impact of DIF: at higher levels of depression and anxiety French patients tended to score a little lower than Dutch patients with the same true levels of depression and anxiety did.

alone in France. Oho-Mpondo suggested that this could hinder the use of the 4DSQ in France. Data collection was laborious and it was suggested that GPs had difficulty in adopting and using this tool in their own practice. Furthermore, the way the questionnaire would be presented to French GPs and their patients would have to be clarified.

Further studies are needed to evaluate the routine use of the 4DSQ in primary care in France, and whether this tool would add value to the consultation. If so, we could follow the Dutch GPs and include 4DSQ to the initial medical training program.

Finally, the validation of the reliability of the 4DSQ in several languages opens the way to further collaborative European studies.

Conclusion

The French 4DSQ measures the same constructs (distress, depression, anxiety and somatization) as the original Dutch 4DSQ, in practically the same way. The French scores can be interpreted in

exactly the same way as the Dutch scores. French GPs' patients filled in the 4DSQ in the same way as Dutch GPs' patients. Accordingly, the French 4DSQ could be used in French General Practice. It could be useful in teaching activities for students and former medical education. It could also be of help for international collaborative studies including French-speaking countries.

Declaration

Acknowledgements

References

1. Shvartzman P, Weiner Z, Vardy D *et al*. Health services utilization by depressive patients identified by the MINI questionnaire in a primary care setting. *Scand J Prim Health Care* 2005; 23: 18–25.
2. Verhaak PF, Hoeymans N, Garssen AA, Westert GP. Mental health in the Dutch population and in general practice: 1987-2001. *Br J Gen Pract* 2005; 55: 770–5.
3. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA* 1999; 282: 1737–44.
4. Terluin B, van Marwijk HWJ, Adèr HJ *et al*. The Four-Dimensional Symptom Questionnaire (4DSQ): a validation study of a multidimensional self-report questionnaire to assess distress, depression, anxiety and somatization. *BMC Psychiatry* 2006; 6: 34.
5. Czachowski S, Terluin B, Izdebski A, Izdebski P. Evaluating the cross-cultural validity of the Polish version of the Four-Dimensional Symptom Questionnaire (4DSQ) using differential item functioning (DIF) analysis. *Fam Pract* 2012; 29: 609–15.
6. Terluin B, Smits N, Miedema B. The English version of the four-dimensional symptom questionnaire (4DSQ) measures the same as the original Dutch questionnaire: a validation study. *Eur J Gen Pract* 2014; 20: 320–6. doi:10.3109/13814788.2014.905826.
7. Van Ginkel J, van der Ark L. SPSS syntax for missing value imputation in test and questionnaire data. *Appl Psychol Meas* 2005; 29: 152–3.
8. Rosseel Y. lavaan: an R package for structural equation modeling. *J Stat Softw* 2012; 48: 1–36.
9. Hu L, Bentler P. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Model* 1999; 6: 1–55.
10. Petersen MA, Groenvold M, Bjorner JB *et al*. Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Qual Life Res Int J Qual Life Asp Treat Care Rehabil* 2003; 12: 373–85.
11. Zumbo B. *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, Ontario,

Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense, 1999.

12. Hambleton RK. Good practices for identifying differential item functioning. *Med Care* 2006; 44: S182–8.

13. Michaelides M. An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Pract Assess Res Eval* 2008; 13: 1–16.

14. Gotzmann A, Bahry L. Review of "jMetrik." *Res Pract Assess* 2012; 7: 56–8.

15. Choi SW, Gibbons LE, Crane PK. lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *J Stat Softw* 2011; 39: 1–30.

16. Violán C, Foguet-Boreu Q, Hermosilla-Pérez E *et al*. Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity. *BMC Public Health* 2013; 13: 251.

17. Nabbe P, Le Reste JY, Robert E *et al*. Family Practice Depression and Multi-morbidity (FPDM): a European consensus on a diagnostic depression tool in primary care. *Eur J Gen Pract* 2013; 19: 162–84.

18. Terluin B, Terluin M, Prince K, van Marwijk HWJ. De Vierdimensionale Klachtenlijst (4DKL) spoort psychische problemen op. [The Four-Dimensional Symptom Questionnaire (4DSQ) detects psychological problems.] *Huisarts Wet* 2008; 51: 251–5 [English translation available on: www.emgo.nl/researchtools/4dsq.asp, accessed on 24 April 2015].

19. Oho Koum-Mpndo J. Exploration des troubles psychosomatiques en médecine générale. Apport de l'index des plaintes en quatre dimensions (IP4D), questionnaire en cours de validation en langue française. Dissertation. Faculté de Médecine, University of Strasbourg, 2013.

20. Schumacher S, Rief W, Brähler E *et al*. Disagreement in doctor's and patient's rating about medically unexplained symptoms and health care use. *Int J Behav Med* 2013; 20: 30–7.