Alice Malpass, Chris Dowrick, Simon Gilbody, Jude Robinson, Nicola Wiles, Larisa Duffy and Glyn Lewis

Usefulness of PHQ-9 in primary care to determine meaningful symptoms of low mood:

a qualitative study

Abstract

Background
Self-administered questionnaires, such as the Patient Health Questionnaire (PHQ-9), are regularly used in clinical practice to guide prescribing or to measure recovery and response to treatment. There are concerns that patients are not all interpreting the questionnaire items in the same way. Cognitive interviewing is a research technique that identifies 'interpretative measurement error' (IME). IME is distinct from traditional components of measurement error, such as not reading the question as worded, or recording answers inaccurately.

Aim
To use cognitive interviewing techniques to explore patterns in answer mapping and comprehension of the PHQ-9 questionnaire to ascertain whether the measure captures meaningful symptoms of low mood.

Design and setting
Qualitative study using cognitive interviewing techniques and card sorting in six GP practices in Bristol.

Method
The study recruited 18 participants at the point of entry to a longitudinal primary care depression cohort study, PANDA (the indications for Prescribing ANtiDepressants that will leAd to a clinical benefit). Participants were interviewed 2, 4, and 6 weeks after their baseline visit. Cognitive interviews were digitally recorded. Analysis used the digital audio file, rather than verbatim transcripts, as it retained important features needed for analyses.

Results
Cognitive interviewing revealed that items on the PHQ-9 are interpreted in a range of ways, that patients often cannot 'fit' their experience into the response options, and therefore often feel the questionnaire is misrepresenting their experience of meaningful symptoms of low mood.

Conclusion
The PHQ-9 may be missing the presence and/or intensity of certain symptoms that are meaningful to patients. Clinicians should adopt caution when using it.

Keywords
depression; cognitive interviewing; PHQ-9; mood questionnaire; prescribing in primary care; safety culture.

INTRODUCTION

Clinicians are encouraged to use mood questionnaires in routine primary care in a range of health settings. In the US, encouragement to use mood questionnaires comes from the US Preventive Services Task Force (USPSTF)[1] and the Agency for Healthcare Research and Quality (AHRQ).[2] In the UK, the Quality and Outcomes Framework (QOF) has encouraged clinician use of brief self-administered questionnaires such as the Patient Health Questionnaire (PHQ- 9) (Table 1).[3] Many GPs do not think the brief severity questionnaires are valid pointers to determine treatment choices[4] and antidepressant prescribing decisions are not based solely on reaching a threshold on the questionnaire.[5] The latest National Institute for Health and Care Excellence (NICE) guideline on depression discourages the sole use of questionnaires to guide prescription.[6] Self-report mental health questionnaires are also increasingly a focus of research.[7]

The qualitative findings presented here are part of a larger study, PANDA (the indications for Prescribing ANtiDepressants that will leAd to a clinical benefit). PANDA is a longitudinal cohort study of people with depression identified in primary care, investigating the clinically important difference on commonly used self-administered questionnaires for depressive symptoms. The PANDA study uses the 'global rating of change' question[8] to estimate a minimal clinically important difference. This approach takes into account the individual's own judgement about whether an improvement has occurred, which can then be compared with the change of scores on questionnaires such as the PHQ-9.

From a cognitive psychology perspective,[9] comparing a global rating of change question with changes in scores on a questionnaire may be problematic because, although self-report measures are validated using standard quantitative approaches, they are not validated for what social theorists call 'interpretative measurement error' (IME):

'The goal of standardisation is that each responder be exposed to the same question experience so that any differences in the answers can be correctly interpreted as reflecting differences between responders rather than differences in the [interpretative and meaning-making] *process that produced the answer.*'[10]

Interpretative differences may be enhanced in patients with depressive symptoms. For example, patients may

A Malpass, PhD, research fellow, Centre for Academic Primary Care, School of Social and Community Medicine; **N Wiles**, PhD, reader in epidemiology, Academic Unit of Psychiatry, Department of Community Based Medicine, University of Bristol, Bristol. **C Dowrick**, MSc, MD, FRCGP, professor of primary medical care, Psychological Sciences, Institute of Psychology Health and Society; **J Robinson**, PhD, professor of the anthropology of health and illness, head of the Department of Sociology, Social Policy and Criminology, School of Law and Social Justice, University of Liverpool, Liverpool. **S Gilbody**, MMedSc, FRCPsych, DipLSHTM, PGDip, professor of primary care mental health, Health Sciences, University of York, York. **L Duffy**, BSc, PANDA trial manager; **G Lewis**, MSc, PhD, MRCPsych, professor of psychiatric epidemiology, Division of Psychiatry, UCL, London.

Address for correspondence
Alice Malpass, Centre for Academic Primary Care, School of Social and Community Medicine, University of Bristol, Bristol BS8 2PS, UK.
E-mail: a.malpass@bristol.ac.uk

Submitted: 15 May 2015; **Editor's response:** 19 June 2015; **final acceptance:** 17 July 2015.

DOI: 10.3399/bjgp16X683473

How this fits in

A handful of studies have used cognitive interviewing with the Beck Depression Inventory. To the authors' knowledge this is the first study to use cognitive interviewing techniques to explore answer mapping and comprehension of the PHQ-9. Research has already shown that clinicians are uncertain about the validity and utility of the PHQ-9 in the management and diagnosis of depression within primary care. This study provides the first empirical evidence that the PHQ-9 may be missing the presence and/or intensity of certain symptoms that are meaningful to patients. As a result clinicians and researchers may want to continue to adopt caution when using and interpreting questionnaire scores with their patients.

struggle more with memory retrieval of relevant information, inhibiting the recall of symptoms over a 2-week period, affecting how they map responses to the options available. Patients may comprehend the same questionnaire item in different ways because of sensitivity towards social desirability, for example, not wishing to disclose suicidal ideation.[11] These are distinct from traditional components of measurement error, such as not reading the question as worded or recording answers inaccurately.[12]

Whereas cognitive psychology is usually interested in process (comprehension and answer mapping), this study also examined the content of responses and their meaning for patients. The main aim was to explore differences between the way patients comprehend and map their answer to the options on the questionnaire. A related aim was to see whether patients shift over time in how they comprehend items on the questionnaire or find them problematic to answer, perhaps in relation to their own changing symptoms.

METHOD

Study design and setting

This was a longitudinal qualitative study design, using cognitive interviewing techniques in six GP practices in Bristol.

Cognitive interviewing

Cognitive interviewing is a method that ensures responders understand questionnaire items in a consistent way, and feel able and willing to provide answers that represent their experience. Unlike most other measures, the PHQ-9 was developed and refined for use with medical patients, not psychiatric patients or community residents. This is important because the criterion validity had to be established in patients with high rates of non-specific physical symptoms that may confound the diagnosis of Major Depressive Disorder.[13] This context helps us evaluate the sort of information its authors intended to generate. Cognitive interviewing is:

'... used to evaluate the quality of response or help determine whether the question is generating the sort of information that its author intended...'[12]

It includes the responders' interpretation of the question and particular terms, their comfort level with answering, any mediating factors that may influence their responses (for example, faith in God or sense of shame), and their own sense of confidence in the accuracy or meaningfulness of their answer (that is, does the box they tick really represent what they feel is the 'truth'?).

Sampling

Participants who had completed baseline data for the PANDA study were recruited into this study. Recruitment took place in Bristol in 2013. A purposive sampling strategy was used. This ensured ethnicity, sex, and sociodemographic differences (using GP practice as a proxy for social demographic of participants) were represented as much as possible. Patients with a range of Clinical Interview Schedule Revised (CIS-R) scores to represent mild, moderate, and severe ICD-10 diagnosis of depression (Table 2) were selected. It was decided in advance to approach 20 participants with the aim of conducting three cognitive interviews with each participant, resulting in 60 interviews for analysis. This size of dataset is large for qualitative research.[14] During analysis a saturation of themes was reached at 18 participants, with 48 completed interviews, and so there was no need to continue recruitment as far as the target of 20.

Data collection

Participants recruited to the PANDA study consented to be contacted about the qualitative study. At the first cognitive interview, participants gave full written informed consent to take part in the qualitative study. Participants were interviewed three times; at 2, 4, and 6 weeks after their baseline appointment. The lead author conducted all the interviews, which lasted between 50 and 180 minutes. Interviews used a protocol guide (summarised below) and were digitally recorded.

Table 1. Scoring on the PHQ-9 Questionnaire

Over the last 2 weeks, how often have you been bothered by the following problems?	Not at all	Several days	More than half the days	Nearly every day
1) Little interest or pleasure in doing things?	0	1	2	3
2) Feeling down, depressed, or hopeless?	0	1	2	3
3) Trouble falling or staying asleep, or sleeping too much?	0	1	2	3
4) Feeling tired or having little energy?	0	1	2	3
5) Poor appetite or overeating?	0	1	2	3
6) Feeling bad about yourself — or that you are a failure or have let yourself or your family down?	0	1	2	3
7) Trouble concentrating on things, such as reading the newspaper or watching television?	0	1	2	3
8) Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual?	0	1	2	3
9) Thoughts that you would be better off dead, or of hurting yourself in some way?	0	1	2	3

The stages of a cognitive interview: the protocol guide. Patients were invited to complete the global rating of change question and the PHQ-9 while thinking aloud what was going through their minds as they read the questions and pondered the answers. The lead author used non-directive, open verbal probing during this process, such as: 'Tell me a bit more about what you are thinking.' Observation probes were used alongside non-directive probing, such as: 'You're hesitating. Can you tell me why?' This was followed up with more targeted probes about the response process, for example, by asking: 'What does that term mean to you?'

Card sorting is an integral part of the cognitive interviewing approach, to determine how individuals organise concepts, in this case, meaningful symptoms.[12] Participants were given a pack of symptom cards, each card having one symptom from the PHQ-9, and asked to rank their symptoms on a scale of 1–10, where 10 represents the most meaningful symptom in terms of impact or intensity, and 1 represents the least meaningful. This prompted a narrative of meaningful symptoms that was digitally recorded. Blank cards were also available for participants to write symptoms that were important to them, which could be placed on the scale.

Data analysis

Digital audio files were used to analyse the data rather than verbatim transcripts because the audio file retains important verbal features needed to contextualise analysis of 'answer mapping' and 'comprehension', such as hesitations and sighing. An Excel grid was created for analysis with 18 column headings, each column heading denoting 'comprehension' or 'answer mapping' for each item on the PHQ-9. Additional columns summarised data from the card sort exercise and the global rating of change question. Participants were listed in rows. For each participant three rows were completed, each row representing a different time point at week 2, 4, or 6. This approach to analysis has similarities to that used in framework analysis.[15]

RESULTS

Of the 20 participants who were approached, two did not respond to initial contact and the remaining 18 were recruited into this study. Of these, 14 completed all three interviews, two participants completed two interviews, and two completed only one interview. In total, 48 cognitive interviews were completed. The age range, CIS-R scores, and GP practice (as a proxy for social demographic) of participants were evenly distributed (Table 2).

The findings explore themes in answer mapping and comprehension using verbatim text from cognitive interviews as illustrations of an issue that, in most cases, affected participants across the sample. Each verbatim quote is tagged with a numerical identifier, the responder's occupation, and whether the data come from the first, second, or third interview. Where appropriate there was referral to the card-sorting data to show under-reporting on the PHQ-9 of a symptom's intensity or impact for the participant. The card-sorting exercise also invited participants to write down their own unique meaningful symptoms on blank cards. Not all patients filled them in. Those that did listed either: perceptual symptoms (improvements in vitality in vision where things look brighter and more vibrant); depersonalisation (where experience slips out of focus); feelings such as resentment, exclusion, and loneliness; and somatic sensations in the body such as tremors, exhaustion, restlessness, a weight on the shoulders, pain in the body, a knot in the stomach, a sense of a ticking time bomb in the body, and nausea. All these symptoms formed a meaningful and/or intense part of their changing low-mood symptoms but were not represented on the PHQ-9. No comprehension or answer-mapping issues

Table 1. Purposive sampling strategy: sample characteristics, *N* = 18

	n
ICD-10 diagnosis: status of depression	
Mild	6
Moderate	6
Severe	6
Sex	
Male	7
Female	11
Ethnicity	
White	9
Black/ethnic or mixed ethnicity	9
Sociodemographic area, name (IMD scores for the wards in which practices are based)	
Hartcliffe (3)	4
Winterbourne (10)	4
Montpelier (3)	4
Clevedon (9)	2
Lawrence Hill (3)	3
Bradley Stoke (3)	1

ICD-10 = International Statistical Classification of Diseases and Related Health Problems, 10th revision. IMD = Index of Multiple Deprivation.

emerged from the global rating of change question.

Participants translated the options on frequency into their own meaningful measure of intensity. For example, 'several days' was used to represent low-level intensity rather than the actual number of days a certain symptom had arisen:

'I feel sad and down sometimes, more than the average person. When I think about things I feel down every day. If I put it nearly every day it would make it look much more severe than it really is. Because I'm not really sure, I'd put several days because it's not committing me. It is every day but only small parts of the day. Especially now I can see more outside of the box, I can stop dwelling on the things that make me low.' (202, GP, 3rd interview)

The same participant wrestled with representing intensity versus frequency of a symptom at more than one interview:

'When it's been there [feeling down, depressed and hopeless] *it's been intense but it's not been as much as more than half the days. It's been intense, but it's not lasted all day. Short lived but more intense.'* (202, GP, 2nd interview)

Similarly, another participant did not answer item 6 (feeling bad about yourself) on the basis of frequency, but on the basis of the intensity and impact of her negative thoughts:

'I'm doing quite well at the moment. I'm going to put "not at all", although there have been episodes of sitting in the car thinking: "Oh God, what a waste of a life — house is a mess, garden is a mess, going to be evicted because you can't pay the rent." Ruminating thoughts have been transitory, they've not settled in on me. I haven't spent that much time really thinking about myself, that nasty churning over.' (181, not working, 2nd interview)

Several triple- or double-barrelled questions caused difficulty. Item 9 (Suicidal ideation) asks if patients have been bothered with 'thoughts that you would be better off dead, or of hurting yourself in some way'. Patients distinguished these two parts of the question as referring to very different things, which made it difficult for them to answer:

'[They are] *different thoughts altogether,* [I'm] *definitely not suicidal, just questioning God: "Why do you keep me alive when there is nothing here for me?" Suicide is self-harm, but I'm asking God: "Why can I not just wake up in the morning, go in my sleep?" Suicidal thoughts at Christmas were completely different feelings, feel as though you not attached to anything, you can drive a car but you don't feel like you, not hooked up to the car, driving it but not part of it, the body felt different.* [Example of a suicidal thought.] *Thinking of driving to the Severn Bridge and jumping off of it. Don't make plans, it's just spontaneous. Thoughts that you would feel better off dead. That doesn't mean self-harm does it? Does that mean suicidal thoughts? It could do, or it could be just wishing you're not here — if so I would put several days, then if it was suicidal thoughts I would put "not at all". If I interpret that as non-suicidal, I'll put several days.'* (162, volunteer at hospital, 1st interview)

Item 6 (feeling bad about yourself, or that you're a failure or have let your family down) also caused problems as participants felt they had experienced different aspects of 'feeling bad' in different frequencies and intensities:

'I do have the bad feelings about myself and those are really intense. I try to minimise the impact on family but I don't know if I always succeed. Certainly the bad feeling about myself has been intense. Do I have it every day? Certainly the bad feeling about myself every day, it's hard because there are three aspects to that. So if was just feeling bad about myself it would be nearly every day, or that you are a failure, more than half the days, or that you've let your family down, probably several days. Feeling bad about myself is a constant and the other feelings are a consequence. I'll tick every day because I can say that.' (172, working mother, 1st interview)

Similarly, another participant could respond to each part of item 6 with different responses:

'That's three different things. If I was answering them separately, feeling down — several days; depressed — more than half the days; hopeless — not at all. I'm not hopeless because I know I can do things. That's three different things. I'd leave that one blank. If I cross the hopeless out, I can answer it.' (188, artist, 2nd interview)

The use of 'or' was confusing, leading participants to wonder: *'Should I answer it if just one applies to me?'* (185), or wanting

to cross out the section that doesn't apply. For example, item 5 (poor appetite or overeating):

'Poor appetite or overeating — it's confusing because it's got both, so I want to cross out overeating, it hasn't affected me, only when I'm depressed, so what do I put?' (182, not working, 1st interview)

Item 7 (concentration) caused comprehension problems because of the specificity of examples, intended to illustrate everyday concentration problems, 'such as watching television or reading the newspaper'. Participants often read this literally:

'That gets me, as it assumes one would normally [watch TV]. *I don't normally do those things. I'd have to be a bit theoretical because I've not watched the television or read the newspaper.'* (202, GP, third interview)

Similarly, other participants also ticked 'not at all' for this item because they do not read newspapers, although they described having trouble concentrating during the card sort exercise.

One participant who was never able to sleep for longer than a few hours each day, found item 3 (trouble falling or staying asleep, or sleeping too much) difficult to understand and misrepresented her experience:

'I'm not getting enough sleep, so not really — "not at all" innit? "Not at all" means not sleeping as much as I am. I would like to sleep longer but I can't, I just automatically wake up. [Researcher probes her comprehension of the item.] *I don't have trouble falling asleep, but I wake up and that's it, I don't go back. So it would be nearly every day.'* (194, cleans trains overnight, 2nd interview)

The findings did not show that patients shift over time in how they comprehend items on the PHQ-9 or find them problematic to answer in relation to their own changing symptoms. On the contrary, the same comprehension and answer-mapping problems were expressed at more than one time point by the same participants, for example, double- or triple-barrelled questions remained problematic over time. However, there was a mismatch between participant perceptions of completing the questionnaire over time in relation to their symptoms. Some participants felt they had completed the questionnaire exactly the same each week because they perceived that their symptoms had not changed, but in practice their responses on the PHQ-9 had changed.

DISCUSSION

Summary

A wide range of comprehension and answer-mapping difficulties were found on the PHQ-9, which persisted over time. Language design issues through the use of double- or triple-barrelled questions were problematic for those who felt they could respond differently to each part of the question. Timescale options were challenging with, for example, a day being experienced as variable. And participants expressed a tension between frequency and intensity of symptoms, also making it difficult for them to map a meaningful answer.

As far as the authors are aware, this is the first study to use cognitive interviewing techniques to explore answer mapping and comprehension of the PHQ-9. The findings demonstrate the value of asking participants what meaning each item on the questionnaire had for them and their reasons for responding to each item as they did.

Strengths and limitations

This study has several limitations. Cognitive interviewing as a methodological approach cannot indicate the size or extent of a problem with particular items on the questionnaire, nor can it guarantee that all problems have been captured, especially as research suggests there is a positive relationship between sample size and problem detection.[16]

Using cognitive interviewing techniques in a longitudinal study design may have led to participants becoming 'schooled' in the questionnaire. The use of 'non-directive' and 'observational probes' during questionnaire completion may have influenced how responders continued to map their answers. However, the findings showed the same issues in comprehension and answer mapping came up at each time point, suggesting participants did not adjust their answers in response to becoming more familiar with the questionnaire, or in response to the interaction of the cognitive interview probes.

Approaches to analysis of cognitive interview data are still being developed and debated.[12] The coded analysis for this study was systematic and drew on the theoretical framework underpinning cognitive interviewing by framing analysis under

'comprehension' and 'answer mapping'.[9] Analysis was not double coded, which is a limitation of the study.

Comparison with existing literature

The problems identified in this study in relation to suicidal ideation items have been reported elsewhere. For example, a comparison of interview data with PHQ-9 responses found patients under-reported suicidal ideation and the measure failed to pick up increases in intensity of suicidal thought that may be less frequent.[11] These findings help explain why this under-reporting is occurring: because of the multiple ways 'thoughts of self-harm' and 'being better off dead' are interpreted as statements.

Another way to view the findings is through the terms adopted by a study interested in the 'discursive fit' between what items demand from informants and what informants decide to do with such a demand.[17,18] The research discusses three strategies informants adopt to cope with problematic items on the Beck Depression Inventory (BDI). They reformulate items, answering different questions from those posed by the questionnaire. They recontextualise items, drawing on contexts that rendered the item nonsensical. Or they contest the assumptions underlying the scale, rejecting it altogether. In the findings reported in this study all three strategies can be seen. For example, item 7 (concentration) was contested by a participant who rejected it as irrelevant because her experience did not match the examples given. Participants also repeatedly contested the meaningfulness of questionnaire items if they were double- or triple-barrelled questions (items 4, 6, and 9). Participants reformulated the options in frequency (not at all, several days, half the days, more than half the days) into their own personalised scale of intensity.

Implications for practice

The findings suggest that the wording on the PHQ-9 could be improved so that patients and clinicians can more usefully distinguish between frequency and severity of symptoms. Research shows that patients who get better while undergoing treatment score better on the PHQ-9, indicating it is a reliable measure of patients' condition and recovery.[3] How do we reconcile psychometric credibility based on quantitative measures of reliability and validity with qualitative analysis, such as this, which raises questions about its use as a measure to represent symptoms that are meaningful to patients?

One plausible explanation is that patients in clinical settings (or research settings) are not encouraged to challenge or comment on the questionnaire, as participants are in cognitive interview studies. They instead routinely engage in 'trying to give the "right" answer', knowing what is at stake[17] and so adopt a 'fake-good profile'.[19] The following commentary on the BDI may equally apply to the PHQ-9:

'The BDI works within the parameters of the dominant discourse of psychiatry and clinical psychology and so it successfully measures something, because it corresponds with the rules of what constitutes such measurement. And while it might identify (Major) Depressive Episode (ICD F32-33 or DSM 296.2-3) it is unlikely to pin down the individual experience of low mood, sadness or what we call "depression".[17]

Patients complete the PHQ-9 in socially situated and power-laden contexts. Researchers stress the importance of qualitative methods in the ongoing evaluation of instruments, to inform quantitative psychometric evaluations and the appropriate use of instruments in clinical practice.[19] The findings are of relevance to ongoing clinical practice because they suggest, as clinicians have suspected for some time, that screening measures are limited when compared to practical wisdom and clinical judgement.[5] Clinicians have expressed uncertainty about the PHQ-9's validity and utility, and in the management and diagnosis of depression within primary care have a strong preference for clinical judgement over scores on severity measures.[5] In light of the numerous ways the PHQ-9 may be missing the presence and/or intensity of certain symptoms that are meaningful to patients, clinicians should continue to adopt caution when using and interpreting questionnaire scores. The study raises the question that longer assessments may be better in providing opportunities for distinguishing frequency and severity, for example, as the CIS-R does.

Funding

This is a summary of independent research funded by the National Institute for Health Research (NIHR)'s Programme Grants for Applied Research Programme (Grant Reference Number RP-PG-0610-10048). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

Ethical approval

Ethical approval was granted by NRES Committee South West — Central Bristol (reference 12/SW/0267).

Provenance

Freely submitted; externally peer reviewed.

Competing interests

The authors have declared no competing interests.

Discuss this article

Contribute and read comments about this article: **bjgp.org/letters**

REFERENCES

1. US Preventive Services Task Force. Screening for depression in adults: US Preventive Services Task Force recommendation statement. *Ann Intern Med* 2009; **151(11):** 784–792.
2. Depression Guideline Panel. *Vol 2. Treatment of major depression. Clinical practice guideline No. 5.* Rockville, MD: US Department of Health and Human Services, Agency for Health Care Policy and Research, 1993. AHCPR Publication No. 93-0551.
3. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001; **16(9):** 606–613.
4. Kendrick T, Dowrick C, McBride A, *et al.* Management of depression in UK general practice in relation to scores on depression severity questionnaires: analysis of medical record data. *BMJ* 2009; **338:** b750.
5. Dowrick C, Leydon GM, McBride A, *et al.* Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: qualitative study. *BMJ* 2009; **338:** b663.
6. National Institute for Health and Care Excellence. *Depression: the treatment and management of depression in adults.* CG90. London: NICE, 2009.
7. Roy T, Lloyd CE. Cultural applicability of screening tools for measuring symptoms of depression. In: Lloyd CE, Pouwer F, Hermanns N, eds. *Screening for depression and other psychological problems in diabetes.* London: Springer, 2013: 67–86.
8. Harmer CJ, Cowen PJ, Goodwin GM. Efficacy markers in depression. *J Psychopharmacol* 2011; **25(9):** 1148–1158.
9. Tourangeau R, Rips L, Rasinski K. *The psychology of survey response.* Cambridge: Cambridge University Press, 2000.
10. Fowler FJ, Mangione TW, eds. *Standardized survey interviewing: minimizing interviewer-related error.* Newbury Park, CA: Sage Publications, 1990.
11. Malpass A, Shaw A, Kessler D, Sharp D. Concordance between PHQ-9 scores and patients' experiences of depression: a mixed methods study. *Br J Gen Pract* 2010; DOI: 10.3399/bjgp10X502119.
12. Willis GB. *Cognitive interviewing: a tool for improving questionnaire design.* London: Sage, 2005.
13. Bombardier CH, Richards JS, Krause JS, *et al.* Symptoms of major depression in people with spinal cord injury: implications for screening. *Arch Phys Med Rehab* 2004; **85(11):** 1749–1756.
14. Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. *Field Methods* 2006; **18(1):** 59–82.
15. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Huberman AM, Miles MB, eds. *The qualitative researcher's companion.* Thousand Oaks, CA: Sage Publications, 2002: 305–329.
16. Blair J, Conrad F, Ackermann AC, Claxton G. The effect of sample size on cognitive interview findings. In: *Proceedings of the American Statistical Association.* Alexandria, VA: ASA, 2006.
17. Galasinski D. Constructions of the self in interaction with the Beck Depression Inventory. *Health* 2008; **12(4):** 515–533.
18. Galasinski D, Kozlowska O. Questionnaire and lived experience: strategies of coping with the quantitative frame. *Qual Inq* 2010; **16(4):** 271–284.
19. Barroso J, Sandelowski M. In the field with the Beck Depression Inventory. *Qual Health Res* 2001; **11(4):** 491–504.