BMJ Open

# Why are there discrepancies between depressed patients' Global Rating of Change and scores on the Patient Health Questionnaire depression module? A qualitative study of primary care in England

Jude Robinson,[1] Naila Khan,[2] Louise Fusco,[2] Alice Malpass,[3] Glyn Lewis,[4] Christopher Dowrick[2]

**To cite:** Robinson J, Khan N, Fusco L, *et al*. Why are there discrepancies between depressed patients' Global Rating of Change and scores on the Patient Health Questionnaire depression module? A qualitative study of primary care in England. *BMJ Open* 2017;**7**:e014519. doi:10.1136/bmjopen-2016-014519

► Prepublication history and additional material are available. To view these files please visit the journal online(http://dx.doi.org/10.1136/10.1136/bmjopen-2016-014519).

Received 29 September 2016
Revised 22 February 2017
Accepted 28 February 2017

[1]Department of Sociology, Social Policy and Criminology, University of Liverpool, Liverpool, UK
[2]Department of Psychological Sciences, University of Liverpool, Liverpool, UK
[3]School of Social and Community Based Medicine, University of Bristol, Bristol, UK
[4]Department of Mental Health Sciences Unit, UCL Psychiatric Epidemiology, London, UK

**Correspondence to**
Prof Christopher Dowrick;
cfd@liv.ac.uk

## ABSTRACT

**Objectives** Our aims were to investigate discrepancies between depressed patients' GlobalRating of Change (GRC) and scores on the Patient Health Questionnaire depression module (PHQ-9). Our objectives were to ascertain patients' views on the source and meaning of mismatches and assess their clinical significance.

**Design** Qualitative study nested within a cohort, in a programme investigating the indications for prescribing antidepressants that will lead to a clinical benefit.

**Setting** Primary care practices in north-west England.

**Participants** We invited 32 adults with a recent diagnosis of depression and evidence of mismatch between GRC and PHQ-9 Scores to participate. Of these, 29 completed our interviews; most were women, identified as white British, had high school education or higher, were employed or retired and had been depressed for a long time.

**Main measures** We conducted semistructured interviews with a topic guide, focusing on experiences of depression; treatment experiences and expectations; effectiveness of the questionnaires; reasons for the mismatch; and social factors. Interviews were transcribed and subjected to interpretative phenomenological analysis.

**Results** We identified four themes as explanations for mismatch between GRC and PHQ-9: perceptions that GRC provided a more accurate assessment of current mental state than PHQ-9; impact of recent negative or positive life events on either measure; personal understanding of depression as normally fluctuating, and tendency to underscore on PHQ-9 as a means of self-motivation; and lack of recall.

**Conclusions** The combined used of the PHQ-9 and a more open question better captures the patient's unique experiences of mental health. This approach ascertains the relevance of symptoms to the individual's experience and influences treatment decisions.

**Study registration** This study was an element of NIHR Programme Grant RP-PG 0610 10048.

## Strengths and limitations of this study

- ► First study to examine mismatch between Patient Health Questionnaire depression module Scores and patients' rating of change.
- ► Graphical presentation of mismatches enabled patients to tell their stories.
- ► Depth narrative approach elicited patients' views and opinions of the relative scores.
- ► Sample was weighted towards patients with chronic depression.
- ► Focus on substantial mismatches may have missed minor fluctuations of significance to patients.

## INTRODUCTION

The Patient Health Questionnaire depression module (PHQ-9) is one of the most widely used self-rating depression scales globally; it is a self-administered instrument consisting of nine items, corresponding to the nine symptoms of depression (principally anhedonia and low mood, together with problems with sleep, appetite, energy, concentration, self-esteem, activity and suicidal ideation) identified in Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV).[1 2] Kroenke *et al*[2] consider the PHQ-9 to be a valid measure to detect and assess depression, which can be used as both a diagnostic algorithm and as a means of measuring the severity of the patient's depressive symptoms.[3] Kendel *et al*[4] find the PHQ-9 to be useful and economical due its brevity and ease of completion.

The Global Rating of Change (GRC) focuses on how the patient is feeling and asks: 'How are you feeling in comparison to two weeks ago?'. It gives the option of five possible answers: 'I feel a lot better', 'I feel slightly better', 'I feel

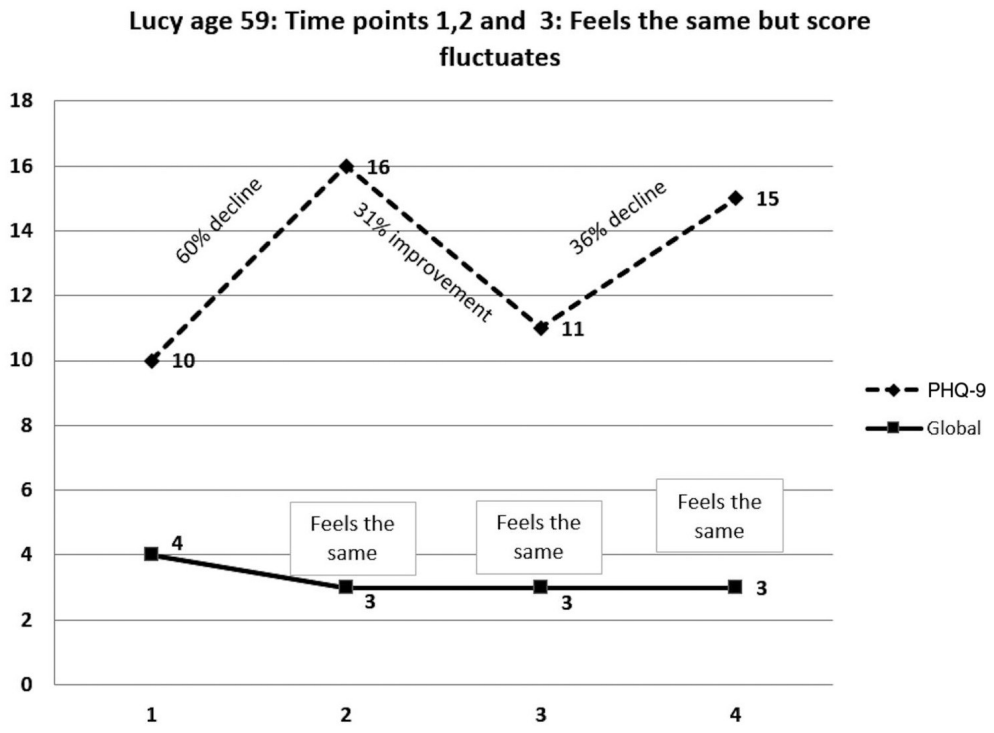


**Figure 1** (Lucy) shows the Global Rating of Change (GRS) Score remaining unchanged at all time points with the participant reporting that they feel the same as they did 2 weeks ago, however their Patient Health Questionnaire depression module (PHQ-9) score is fluctuating substantially across the time points. This figure displays examples of mismatches between the PHQ-9 Score and the GRC. The x axis represents the four biweekly assessments and the y axis represents the score. The dotted line represents the PHQ-9, where an increased score indicates a decline in symptoms. The solid line represents the GRC, again an increased score indicates a decline (1=feels much better, 2 = feels better, 3= feels the same, 4 = feels worse, 5= feels much worse).

about the same', 'I feel slightly worse' or 'I feel a lot worse'. The GRC is used within both research and clinical contexts and can be used in clinical trials research as an outcome measure alongside observer-report or self-report questionnaires.[5 6] It has the advantage of allowing patients to take into account factors they consider important to their particular situation, but the disadvantage that the assessor may not know what the patient is taking into account when making their rating.

Research in general practice suggests that general practitioners prefer to diagnose and manage depression according to their clinical knowledge, although clinical interpretations of depression do not always reflect the patient's personal experience.[7] While patients consider rating instruments of depression to be both useful and informative, such measures may not adequately reflect the reality of patient's experiences and recovery.[8–13] It may be that self-rated instruments and patient GRCs provide complementary perspectives of value in clinical encounters, but there is little existing research to support this assumption.

The aim of this study is to investigate discrepancies between depressed patients' GRC and their PHQ-9 Scores. The objectives are to ascertain patients' views on the source and meaning of such 'mismatches', and assess their clinical significance.

## METHODS

### Selection of participants

The study sample was drawn from a cohort study within the PANDA programme (NIHR programme "What are the indications for Prescribing ANtiDepressAnts that will lead to a clinical benefit?") investigating the indications for prescribing antidepressants that will lead to a clinical benefit (NIHR Programme Grant – RP PG 0610 10048). This element of the programme was an initial naturalistic study, undertaken prior to a subsequent placebo-controlled antidepressant drug trial, and was not related to trials of any specific interventions.

The cohort sample was generated through electronic searches conducted by participating primary care practices to identify potential participants aged between 18 years and 70 years and diagnosed with depression within the last 12 months, excluding patients who had comorbidity with bipolar disorder, eating disorder or psychosis, a major alcohol or substance abuse problems or were 30 or more weeks pregnant. During the course of quantitative interviews conducted by researchers, participants in the PANDA cohort study completed the PHQ-9 and the GRC at four time points: at baseline and at week 2, week 4 and week 6 of the study.

We examined the data of 86 individuals completing all cohort study assessments and identified that the results

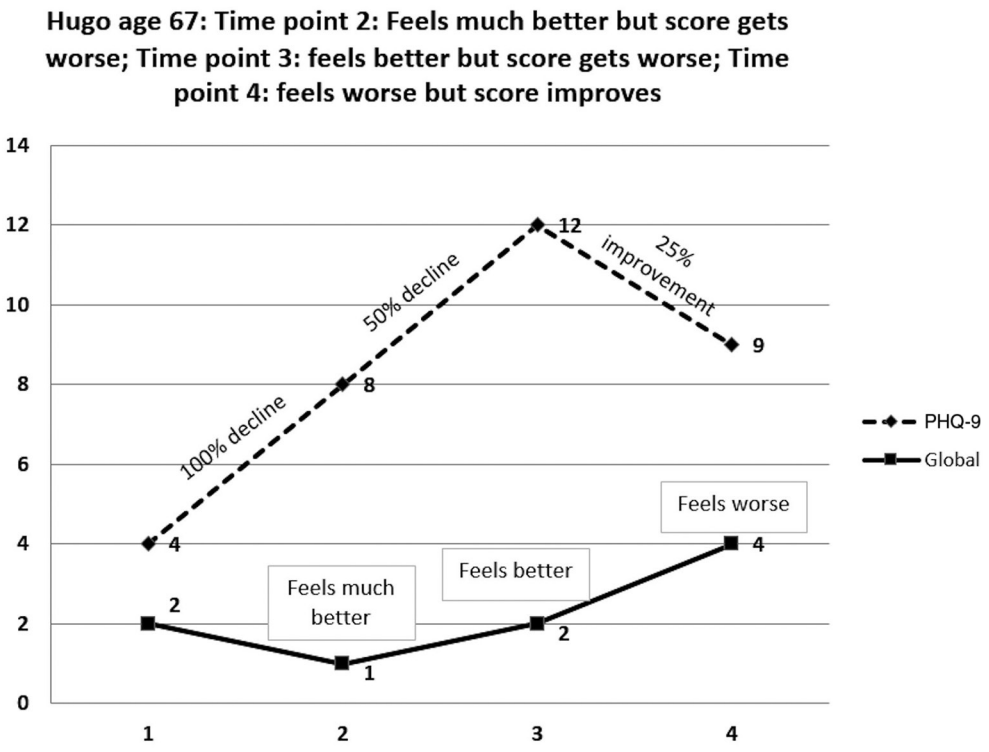


**Figure 2** (Hugo) shows a mismatch where the global rating and the Patient Health Questionnaire depression module (PHQ-9) score are entirely incongruent. Where the PHQ-9 Score is declining at weeks 2 and 3 the participant reports that they feel much better and better, respectively, where the PHQ-9 Score improves at week 4 the participant reports that they feel worse. This figure displays examples of mismatches between the PHQ-9 Score and the Global Rating of Change (GRC). The x axis represents the four biweekly assessments and the y axis represents the score. The dotted line represents the PHQ-9, where an increased score indicates a decline in symptoms. The solid line represents the GRC, again an increased score indicates a decline (1=feels much better, 2 = feels better, 3= feels the same, 4 = feels worse, 5= feels much worse).

from the PHQ-9 and the GRC at the different time points appeared not to be in alignment, that is, they were 'mismatched'. Given recent evidence that minimum clinically important difference as perceived by patients is dependent on baseline severity, and that it is best measured on a ratio or percentage scale rather than in absolute terms,[14] we identified a symptom score change of 15% or more in improvement or decline, combined with a GRC Score that indicated either no change or a change in the opposite direction, as a clinically important mismatch. On this basis we identified 44 (51%) cases of mismatch between PHQ-9 and global rating (see figures 1 and 2). The 32 individuals with the most pronounced mismatches were invited to participate by letter, with subsequent telephone reminder if needed. None refused to participate and 29 people completed their interviews. We reached data saturation at this point and therefore did not seek further participation. All of the participants had an established relationship with either LF or NK having completed the initial four assessments. As a product of taking part in the initial assessments the participants had a basic understanding of the researcher's personal research interests and reasons for conducting the interviews.

Our interview sample comprised 10 men and 19 women with an average age of 52 years, ranging from 24 to 68 years (see table 1). All participants were Caucasian, with over 80% identifying as white British. Only 17% of the sample was unemployed, the majority were either employed (31%) or retired (24%). Five individuals (17%) were registered as permanently sick or disabled, but all had previously been employed. Three (10%) participants were engaged with the full time running of home and family. The educational levels were high with more than 69% of the sample having a high school education or higher and more than a quarter of the sample holding a university degree or higher degree.

## Data collection

The qualitative interviews were conducted during September and October 2014, following the completion of a set of four quantitative interviews (including PHQ scoring on each occasion) with each participant during the preceding 16 months. The range of time between completing the quantitative interviews and undertaking the qualitative interview was between 6 weeks and 14 months.

Interviews were conducted by two female postdoctoral researchers NK (registered counsellor) and LF (chartered psychologist), using an interview guide that centred on five key topics: experiences of depression; treatment experiences and expectations; effectiveness of the questionnaires; reasons for the mismatch; and social factors

Table 1 Description of participants

| Participant | Age/gender | Reason for mismatch |
|---|---|---|
| **Jerry** | 49/M | GRC—Felt worse each week but PHQ-9 Score remained the same. |
| **Hugo** | 67/M | GRC—Felt the same at week 2 but PHQ-9 Score remained the same. |
| **Laura** | 36/F | GRC—Felt much worse at week 2 and 3 but PHQ-9 Score improved. |
| **Michael** | 56/M | GRC—Felt slightly better at week 2 but PHQ-9 Score declined. Felt the same at week 3 but score declined. |
| **Gary** | 51/M | GRC—Felt the same throughout, PHQ-9 Score improved at week 2 and week 4. |
| **Jenny** | 46/F | GRC—Felt the same at weeks 2 and 3 but declined in week 4 and then improved. Felt slightly better at week 4 but PHQ-9 Score stayed the same. |
| **Donna** | 45/F | GRC—Felt the same at week 2 but then the score declined. PHQ-9—Felt slightly better at week 4 but the score declined. |
| **Brenda** | 61/F | GRC—Felt the same throughout but PHQ-9 Score fluctuated. |
| **Sid** | 67/M | GRC—Felt slightly better at week 3 but later scores declined. PHQ-9—felt the same at week 4 but score improved. |
| **Brian** | 53/M | GRC—Felt slightly worse at week 3 but PHQ-9 Score remained the same. |
| **Susan** | 53/F | GRC—Felt the same at week 3 but score declined. PHQ-9—felt the same at week 4 but score improved. |
| **Freya** | 26/F | GRC—Felt slightly worse at week 2 but score improved. PHQ-9—Felt slightly worse at week 3 but score declined. Felt the same at week 4 and score remained the same. |
| **Toby** | 48/M | GR—Felt the same at week 2 but the score declined. PHQ-9 —Felt much worse at week 4 but score improved. |
| **Penny** | 54/F | GRC—Felt the same throughout but PHQ-9 Score fluctuated. |
| **Elizabeth** | 50/F | GRC—Felt the same at week 2 but PHQ-9 Score improved. |
| **Olive** | 24/F | GRC—Felt the same at week 3 but PHQ-9 Score improved. GRC—Slight slightly worse at week 4 but PHQ-9 Score dramatically declined. |
| **Maddy** | 43/F | GRC—Felt the same at week 2 but PHQ-9 Score improved. GRC—Felt better at week 3 but PHQ-9 Score remained the same. GRC—Felt the same at week 4 but PHQ-9 Score improved. |
| **Tom** | 65/M | GRC—Felt the same at week 3 but PHQ-9 Score declined. GRC—Felt worse at week 3 but PHQ-9 Score improved. GRC—Felt the same at week 4 but PHQ-9 Score declined. |
| **Stacey** | 62/F | GRC—Felt much worse at weeks 2 and 3 but PHQ-9 Score stayed the same and then improved. Felt much better at week 4 but PHQ-9 Score stayed the same. |
| **Lucy** | 59/F | GRC—Felt the same throughout but PHQ-9 Score fluctuated. |
| **Sophie** | 55/F | GRC—Felt the same at week 2 but PHQ-9 Score declined. GRC—Felt much worse at week 3 but PHQ-9 Score improved. |
| **Anne** | 51/F | GRC—Felt much worse at week 2 but PHQ-9 Score improved. GRC—Felt much worse at week 4 but PHQ-9 Score remained the same. |
| **Pippa** | 59/F | GRC—Felt the same throughout but PHQ-9 Score fluctuated. |
| **Vivien** | 35/F | GRC—Felt the same at week 3 but PHQ-9 Score declined. |
| **Alice** | 66/F | GRC—Felt the same throughout but PHQ-9 Score fluctuated. |
| **Ryan** | 57/M | GRC—Felt the same throughout but PHQ-9 Score fluctuated. |
| **Pauline** | 52/F | GRC—Felt much worse at week 3 but PHQ-9 Score stayed the same. |
| **Sally** | 59/F | GRC—Felt the same at week 3 but PHQ-9 Score declined. Felt much worse at week 4 but PHQ-9 Score remained the same. |
| **Jim** | 66/M | GRC—Felt worse at weeks 2 and 3 but PHQ-9 Score remained the same. Felt the same at week 4 but PHQ-9 Score declined. |

(see online supplementary appendix 1). This guide was designed to elicit a rich narrative of their experience of depression and to contextualise the mismatch discussion. In addition, a plot showing the participants' scores and mismatches was used as a visual guide for each participant during the interviews (see figures 1,2). Their completed questionnaires were also shown to the participants during the discussion of the mismatch, focusing on PHQ-9 and GRC. All the interviews took place on University of Liverpool premises and ranged between 40 min to 2 hours in

length. No one else was present during the interviews, and field notes were not routinely taken. All the interviews were digitally recorded and transcribed. Transcripts were not returned to the participants for comment or correction and participants were not asked to provide feedback on themes.

### Data analysis

We used Interpretative Phenomenological Analysis (IPA) to guide the analysis as this enabled us to focus on the individual accounts before moving to identify more general themes in the data.[15] Four members of the team separately coded and then discussed the same three selected transcripts to enhance inter-rater reliability, and before LF and NK coded all the transcripts to identify initial themes that were then further analysed to formulate superordinate themes and subthemes.[15 16] This approach emphasises the meaning-making nature of the process and is reflected in participants' accounts. It allows researchers to establish a deeper understanding of how people construct knowledge and meaning of their lived experiences.[17–19] Interviews and analyses were conducted concurrently; emergent themes and data saturation were discussed by the team on an ongoing basis.

## RESULTS

Twenty-five (86%) participants reported experiencing depression for more than 10 years, with 18 (62%) identifying as having been depressed since adolescence. The other four (14%) participants reported their first experience of depression within the last 2 years. The mismatch explanations were offered as part of a wider narrative exploring the participants' experiences of depression. We identified four superordinate themes arising from the analysis, with the majority of participants identifying between two and four of these as explanations for their particular mismatch, namely: problems with the PHQ-9 measure; negative and positive life events; personal understanding of depression and coping mechanisms; and, an inability to recall the reason for a possible mismatch. These four themes are explored in more detail below. All participants have been allocated pseudonyms and we have indicated their ages.

### Problems with the PHQ-9

Thirteen participants offered mismatch explanations that centred on perceived shortcomings in thePHQ-9. Several felt unable to engage and respond to the PHQ-9 items, resulting in inaccurate answers and a mismatched result. They felt able to answer the GRC question with a greater degree of accuracy and that this was a truer representation of the state of their mental health.

Ann described how she had felt much worse in week 2 but her PHQ-9 Score showed an improvement, and when she felt even worse in week 4, her score remained the same:

> *To be honest with you this is what was more important to me the way I feel I've wrote down… if it doesn't coincide with that… I know what I feel in my mind there but I just find it a bit difficult explaining it with a tick list […] You know I'd rather have said black and white, 'are you depressed are you not depressed?' (Anne, aged 51 years)*

Several participants thought that not all relevant depressive symptoms were covered by thePHQ-9. It did not allow them to adequately express changes in their symptoms. Missing items include the tendency to withdraw from people, lack of libido and the sudden onset of an inability to cope at work.

> *Erm, there isn't a lot about interacting with other people which I think is the first thing that goes for me … I you know I shut off a bit (Laura, aged 36 years)*

Others commented that they found the PHQ-9 internally repetitive. One participant commented that he found some items to be obsolete, in that low mood is defining of depression and how can something that is defining of depression differentiate between people experiencing depression.

> *You see those last two asking if you've been bothered with problems and then insist on doing things you ask someone on depression if they've got little interest or pleasure doing things you'll probably find that most people say nearly every day on that one… feeling down depressed or restless you know to me it's not insulting but it's like well that's what I'm here for (Jerry, aged 49 years)*

Other participants felt that inadequacies in the PHQ-9 scoring system were the cause of the mismatch. They thought the scale was too crude and did not provide opportunity to accurately express the extent of their symptoms. For example Jerry, scored highly on the PHQ-9, in the following weeks he reported that he was feeling much worse however his PHQ-9 Score remained the same. The reason for the mismatch was that the initial high score did not provide him with opportunity to score any higher on the PHQ-9 Scale:

> *Well I don't know I mean 'not at all', 'several days more' or 'not often' maybe you should do a 1 to 10 thing? I think it would be better you know then they can write in 5/10 that will give them because sometimes, it's just in-between them lines you know? It's not like several days, more or half the days, nearly every day.…You know there's times I've though well it's not that but it's not that either you know. More than half the days or nearly every day well what would that be that would be a 6 and that a 10, generally its 7 or an 8 (Jerry aged 49 years)*

For eight of the participants their depressive symptoms were secondary to other, more prominent symptoms (for example anxiety, PTSD symptoms or physical illness) which determined how they answered the global rating question. So for these participants, the mismatches between the GRC and PHQ-9 Scores arose when they

experienced fluctuations in the symptoms for their *other condition* rather than their depression, but the changes they experienced were not reflected in their PHQ-9 Score as this measured depressive symptoms only.

Gary developed depression 10 years ago after suffering a stroke, resulting in reduced mobility and the termination of his job as a bus driver. For him depression was an afterthought to his ailing physical health, which remained poor. He was able to reflect this on the GRC by saying that he felt the same in terms of his physical health across all weeks. The mismatch arose because the PHQ-9 indicated an improvement in his mental health only:

> *It's more the health, physical well-being you know rather than mental well-being? Well the physical, if the physical side of you is not right, then then the mental side is not going to be right is it?* […] *scores show otherwise and that was the reason why because of the way the health, my health you know physical side of things weren't changing (Gary, aged 51 years)*

### Negative and positive life events

Nineteen participants talked about life events that occurred around the time of the mismatch which had a significant bearing on how they were feeling. Mismatches occurred where life events had a more immediate effect on one of the measures than the other.

Brian, who felt slightly worse week 3 but scored the same, described how he experienced a delayed reaction to stress, resulting in mismatches in his scores. He explained that when he experiences a stress event it is only in the days and weeks to follow that he feels overtly stressed. He explained that 'in the moment' he is consumed by dealing with the situation at hand, he then experiences a 'fall out' where the stress catches up with him and he feels worse:

> *… but I might not of felt it at the time… that might have had quite a big toll on me having to deal with it, I would describe that as a kickback… but because it was for my immediate family I overrode the anxiety and the anxiety kept growing yeah, like I said, before I can feel it ratcheting up (Brian, aged 53 years)*

Overall, 17 participants clearly identified negative life events during the time the mismatch occurred:

> *…I've had, erm, some problems with my daughter she had a baby 18 months ago who was born with a serious heart defect… I came back from where I was living in Spain to try and help her out with that but there became a point where I had to go back… and unfortunately she felt that I'd let her down by going back so our relationship has been very strained (Susan, aged 53 years)*

A few participants attributed their mismatches to positive life events or a newly acquired protective factor. These included being offered extra support at work and managing to access help through counselling or support groups:

> *Probably from a work point of view I was in a very different place then because I had somebody come in to sort of give me a lift with, erm, you know work []…. and I felt such a difference in that because I'd felt quite sort of dragged down by that and that gave me you know sort of boost… (Maddy, aged 43 years)*

Some participants talked about the natural trajectory of their illness, and attributed the changeable nature of depression to their mismatched results. Being 'up and down' was perceived as 'normal', and so fluctuations in symptoms were not necessarily perceived as 'improvements' or 'declines' but as the natural rhythm of their depression. Therefore no change was perceived by the participant.

> *'So it's not an issue for me that even if I felt slightly worse or slightly felt better I think saying I feel about the same is me just saying well yeah this is my normal'(Jenny, aged 46 years)*

### Personal understanding of depression and coping mechanisms

Six participants explained that they had deliberately underscored either the PHQ-9 or the GRC in an attempt to gain control over their depression, with the hope that their symptoms would improve as a result of their positive outlook. Three women made specific statements about underscoring the PHQ-9, of which two are given here:

> *Psychologically I was forcing myself to you know I'm affirming to myself I'm feeling fine… but really I wasn't… I was reading these things about affirmations you know changing your negatives into positives I was exercising that (Olive, aged 24 years)*
>
> *It's like I wanted to convince myself of getting better so I could probably react better… I felt I had to prove myself that… I could help myself better (Elizabeth, aged 50 years)*

Some participants explained that the mismatch had occurred because they did not want to admit how they were feeling to themselves or others when completing the measures. Participants had different motivations for attempting to hide their feelings. Jenny describes the fear of admitting she was feeling worse:

> *… and then I don't know whether there's a bit of that not wanting to its like admitting it if you're ticking those boxes, I think that's the one bit of the test I really, really struggle with… cause that's been my defence mechanism all these years… It's easier to say the same and I think with these questionnaires there is a fear of actually admitting that you feel worse(Jenny, aged 46 years)*

Laura explained that she deliberately omitted item 9 (suicide intention) on the PHQ-9 in order to prevent intervention from a crisis team. As a result her overall PHQ-9 Score appeared reduced where she reported she was actually feeling worse:

*… the second one the one where it looks like I'm only a bit worse I did underscore that because I didn't want you to write to my GP… and the other one that's in here is that I never answer that one honestly ever 'cause that's really going to end you up in hospital so I never answer anything asking if you think you're better off dead or wanting to hurt yourself(Laura, aged 36 years)*

In common with other participants, she also omitted or underscored on items that she found distressing or particularly upsetting to answer truthfully:

*I didn't answer that one* [Item 6: Feeling bad about yourself, or that you are a failure or have let yourself or your family down]… *I just felt so bad… yeah because I do feel like I'm letting Phil down…. erm just about when we got married I was fit and healthy and not the same person I am now. You know not only is there depression now on top of that is chronic physical illness which is not really what any of us signed up for so yeah (Laura, aged 36 years)*

### Inability to remember

Six participants described an inability to recall what was happening at the time the measures were taken or how they were feeling at the time. They observed that they generally found it difficult to recall what they were doing or how they were feeling from 1 day to the next:

*… I can't remember sometimes what I felt like yesterday, so sometimes it is easier to just say the same… how I felt, partly because I couldn't remember how I felt two weeks ago, I can't remember how I felt two days ago sometimes(Jenny, aged 46 years)*

*… I can't remember what I had for my breakfast do you know what I mean?(Jerry, aged 49 years)*

Therefore being asked to recall how they were feeling when the measures were taken during the previous quantitative interviews, and to then make a meaningful comparison to their mood state at that time, was impossible for some participants.

We did not find evidence elsewhere in these interviews to disconfirm or contradict these four superordinate themes.

## DISCUSSION

### Summary

Perceived problems with the PHQ-9 measure played a central role in the mismatch, as some participants were unable to engage with the PHQ-9, felt there were missing or obsolete items, and thought the GRC was preferable as it allowed them to 'sum up' how they were feeling. Participants also took issue with the discrete scoring system of the PHQ-9, suggesting that it was crude and did not allow them enough scope to adequately express the intensity of their symptoms. The GRC also enabled others to reflect on any comorbid conditions, as their depressive symptoms were a secondary consideration to their anxiety or physical symptoms, and this appeared to be the underlying cause of the mismatches we detected. These findings support the validity of using the GRC to determine minimal clinical important difference, as the only criticism of the GRC was that some found it difficult to recall how they'd been feeling retrospectively.

Approximately two-thirds of participants linked life events to their mismatches. They reported that the impact of such events was not immediately felt so did not immediately impact on their assessment of the severity of their symptoms for depression, but affected them in other ways that they summarised in the GRC. For many participants fluctuations in their symptoms did not register as a change as 'ups and downs' are an established part of the natural trajectory of their disorder. Some participants were very clear about their conscious efforts to reduce their symptoms through positive affirmations and so under-reported their symptoms on the PHQ-9. Similarly, others acknowledged reticence at admitting how they were feeling and again underscored on the PHQ-9. This suggests that special care needs to be paid to 'omitted' questions as there is likely to be an explanation for this apparent 'gap'. The final theme 'cannot remember' highlights the difficulty participants experienced in recall and calls into question the effectiveness of any assessment that asks people to retrospectively assess their mental health over longer time periods.

### Limitations

These findings are limited by the relatively small sample but, as the average age for participants was 56 years, many were retired, chronically depressed, and over two-thirds were educated to secondary school level or above, and 11 held a university degree or higher, they may not be representative of the wider population.

However this may also confer some advantages. Our observations are that people with longer-term depression have further insight and more detailed narratives in terms of possible causes, coping mechanisms and seeing depression as part of their life story. They tend to recognise the changeability of their symptoms over time and are more skilled at recognising in themselves when they were starting to feel worse. We did not find different perspectives between those who were feeling better and those who were feeling worse; rather, the changeability was seen as all 'par for the course'.

The participants were selected because they had completed all the weeks of the first phase of the wider PANDA study and so may be more motivated than others. In addition, our decision to identify a 'mismatch' as more than a 15% fluctuation between the PHQ-9 and GRC Scores may have overlooked minor fluctuations of less than 15% but that were significant to the patients.

It is possible that the passage of time between the PHQ scoring and the qualitative interviews, and the use of IPA methodology, could have led participants to rationalise inconsistencies which were due to measurement error, or to construct retrospective quest narratives. However the depth and detail of participant explanations argue

against discrepancies being due to plain measurement error. Most respondents—apart from those in the 'can't remember' category—had a clear grasp on why there was a mismatch, and there was no evidence of 'distance decay'. The use of IPA methodology did not appear to be a problem as many participants embraced the opportunity to discuss and clarify the differences between self-rating and PHQ-9 Scores, and were clear in their description of what was happening for them at the particular time of the interview. Some participants were aware of the mismatch prior to the discussion with the interview, while others were surprised by their inconsistency across the two measures. In either case, participants were keen to explain why this might have been, and found the IPA methodology valuable in enabling them to do so.

### Strengths

We are not aware of any other studies that have produced evidence from patients with depression who report a mismatch between PHQ-9 Scores and their own GRC. The question of mismatched results and why they occur is relatively complex to frame and communicate to participants. The meaningful research findings gauged from this study would suggest that the design and methods applied were successful and appropriate. Participants were invited to talk about their mismatch to tell their story of depression in a safe and non-judgemental environment. The presentation of mismatches in graphical form aided this process. It is likely that this holistic approach to answering the research question resulted in the rich narrative surrounding the rather constrained question of mismatched results.

### Comparison to existing literature

These findings contrast with those of previous research, and suggest that the PHQ-9 is neither exhaustive in its scope nor transparent or straightforward to complete and may miss symptoms that are meaningful to patients and underestimate their intensity.[1 3 4 20] Approximately two-thirds of the sample linked life events to their mismatches, with stressful events causing or maintaining depression and positive events and support acting as protective factors against relapse and as aids to recovery.[12 21]

Some participants were clear about their conscious efforts to under-report their symptoms in an effort to reduce their symptoms through positive affirmations. This novel finding is likely to resonate with populations engaged with cognitive behavioural therapy and encouraged to 'reframe' their depression.[22] It extends the small existing literature on 'gaming' and bias in relation to self-rating instruments for depression,[8 12 23] and addresses the previously recognised need to consider response bias in mental health research.[24]

Less common than other themes, but asserted with clarity and certainty, was participants' reticence at admitting how they were feeling both to themselves and others. The final theme 'cannot remember', while a common phenomenon, may reflect the impact of depression on memory and recall capacity that is well documented in the wider depression literature.[25]

### Implications for theory and practice

These findings indicate that the processes and motivations behind completing the PHQ-9 are complex, with responses influenced by ongoing physical, social and emotional issues. Furthermore the majority of the sample offered a number of explanations to explain their mismatch suggesting that the completion process is complex and multifaceted. The PHQ-9 is commonly used as a diagnostic tool, where a score above a prescribed threshold is seen as indicative of clinical depression. In line with National Institute for Health and Care Excellence and Quality and Outcomes Framework (QOF) guidelines, our findings emphasise that the PHQ-9 should not be used as a standalone tool but as a diagnostic aid: the role of the PHQ-9 is to *assist* and not *determine* the diagnosis and management of depression, within the context of the clinical holistic encounter between patient and practitioner.[26 27] We note that our findings may be generalisable to other self-rating instruments, though we cannot assume that they are.

Although the PHQ-9 is a self-report measure, its content is constrained by its authors and the clinicians who use it. Use of the PHQ-9 in conjunction with an open-ended enquiry, such as the GRC, would offer patients the opportunity to express themselves outside of the parameters of the measure. A global rating such as 'how are you feeling in yourself?' has historically been used as an opening question in consultation.[28] However it is an approach that is falling out of favour in a wave of over-reliance on quick, efficient psychometric measures such as the PHQ-9. In addition global ratings, although still used to facilitate initial discussion, are less commonly used in follow-up sessions to ascertain change. We consider that questions on patients' GRC may be opportune in all consultations. A careful unpacking of patient's symptoms, and investment in patient's individual experiences, is at the forefront of a productive doctor-patient relationship. Establishing and nurturing that relationship is paramount to the long-term mental health and well-being of the patient throughout their journey to manage their condition.

This research suggests that the combined use of the PHQ-9 and a more open, explorative question would better capture the patient's unique ongoing experiences of mental health. This approach would ascertain the breadth and severity of symptoms and their relevance to the individual's experience, and this would influence treatment decisions.

**Acknowledgements** The authors thank Larisa Duffy for her support of the PANDA study and Maxine Martin and Pam Galloway for their support in recruitment and interviews.

**Contributor** JR and AM conceived and designed the study with CD and GL. LF and NK conducted the interviews and qualitative analysis, and drafted the manuscript. JR, CD, GL and AM critically revised the manuscript. All authors approved the final manuscript; I, CD, affirm that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted.

**Funding** This study was funded by the NIHR Programme Grants Scheme, (NIHR Programme Grant – RP PG 0610 10048). However the researchers remain independent from the funders and the funders did not play a role in the study design or data collection for the study.

**Competing interests** None declared.

**Patient consent** Not obtained.

**Ethics approval** This study was granted ethical approval by Bristol Research Ethics Committee Centre (12/SW/0267).

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data available.



## REFERENCES

1. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001;16:606–13.
2. Kroenke K, Spitzer RL, Williams JB, *et al*. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *Gen Hosp Psychiatry* 2010;32:345–59.
3. Forkmann T, Gauggel S, Spangenberg L, *et al*. Dimensional assessment of depressive severity in the elderly general population: psychometric evaluation of the PHQ-9 using rasch analysis. *J Affect Disord* 2013;148:323–30.
4. Kendel F, Wirtz M, Dunkel A, *et al*. Screening for depression: rasch analysis of the dimensional structure of the PHQ-9 and the HADS-D. *J Affect Disord* 2010;122:241–6.
5. Lin CH, Lu MJ, Wong J, *et al*. Comparison of physician-rating and self-rating scales for patients with major depressive disorder. *J Clin Psychopharmacol* 2014;34:716–21.
6. Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. *J Man Manip Ther* 2009;17:163–70.
7. Ridge D, Ziebland S. Understanding depression through a 'coming out' framework. *Sociol Health Illn* 2012;34:730–45.
8. 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.
9. Leydon GM, Dowrick CF, McBride AS, *et al*; QOF Depression Study Team. Questionnaire severity measures for depression: a threat to the doctor-patient relationship? *Br J Gen Pract* 2011;61:117–23.
10. Mitchell C, Dwyer R, Hagan T, *et al*. Impact of the QOF and the NICE guideline in the diagnosis and management of depression: a qualitative study. *Br J Gen Pract* 2011;61:279–89.
11. Shaw EJ, Sutcliffe D, Lacey T, *et al*. Assessing depression severity using the UK quality and outcomes framework depression indicators: a systematic review. *Br J Gen Pract* 2013;63:309–17.
12. Malpass A, Shaw A, Kessler D, *et al*. Concordance between PHQ-9 scores and patients' experiences of depression: a mixed methods study. *Br J Gen Pract* 2010;60:231–8.
13. Keeley RD, West DR, Tutt B, *et al*. A qualitative comparison of primary care clinicians' and their patients' perspectives on achieving depression care: implications for improving outcomes. *BMC Fam Pract* 2014;15:13.
14. Button KS, Kounali D, Thomas L, *et al*. Minimal clinically important difference on the beck depression inventory--II according to the patient's perspective. *Psychol Med* 2015;45:3269–79.
15. Smith JA, Flowers P, Larkin M. *Interpretative phenomenological analysis: theory, method and research*. London: Sage, 2009.
16. Smith JA, Jarman M, Osborn M. Doing Interpretative Phenomenology Analysis. In: Murray M, Chamberlain K, eds. *Qualitative health psychology: theories & methods*. London: Sage, 1999.
17. Bourdieu P. *Outline of a theory of practice*. Cambridge: Cambridge University Press, 1977.
18. Burr V. *Social constructionism*. 2nd ed. London: Routledge, 2003.
19. Reid K, Flowers P, Larkin M. Exploring lived experience. *The Psychologist* 2005;18:20–3.
20. Malpass A, Dowrick C, Gilbody S, *et al*. Usefulness of PHQ-9 in primary care to determine meaningful symptoms of low mood: a qualitative study. *Br J Gen Pract* 2016;66:e78–e84.
21. Brown G, Harris T. *Social origins of depression*. London: Tavistock, 1978.
22. Beck J. *Cognitive behavior therapy: basics and beyond*. New York: Guildford, 2011.
23. Hunt M, Auriemma J, Cashaw AC. Self-report bias and underreporting of depression on the BDI-II. *J Pers Assess* 2003;80:26–30.
24. Rogler LH, Mroczek DK, Fellows M, *et al*. The neglect of response bias in mental health research. *J Nerv Ment Dis* 2001;189:182–7.
25. Trivedi MH, Greer TL. Cognitive dysfunction in unipolar depression: implications for treatment. *J Affect Disord* 2014;152-154:19–27.
26. NICE *depression in adults: the treatment & management of depression in adults*.*NICE Clinical Guideline*. London; British Psychological Society, 2009;90.
27. BMA & NHS Employers. *Quality and outcomes framework for 2012/3*. London: British Medical Association, 2012.
28. Neighbour R. *The inner consultation*. 2nd Edition. London: Radcliffe, 2005.