Acta Psychiatr Scand 1994: 90: 446–450
Printed in Belgium – all rights reserved

ACTA PSYCHIATRICA SCANDINAVICA
ISSN 0001-690X

Temporal variation of validity of self-rating questionnaires: repeated use of the General Health Questionnaire and Zung's Self-rating Depression Scale among women during antenatal and postnatal periods

Kitamura T, Shima S, Sugawara M, Toda MA. Temporal variation of validity of self-rating questionnaires: repeated use of the General Health Questionnaire and Zung's Self-rating Depression Scale among women during antenatal and postnatal periods.
Acta Psychiatr Scand 1994: 90: 446–450. © Munksgaard 1994.

The 30-item General Health Questionnaire (GHQ) and Zung's Self-Rating Depression Scale (SDS) were distributed to 120 pregnant women 4 times – in early and late pregnancy and 5 days and 1 month after the child was born. The validity of the questionnaires was assessed against the subjects' Research Diagnostic Criteria (RDC) diagnoses. Both the GHQ and SDS sufficiently identified cases of minor mental disorder and depressive disorders respectively in early pregnancy; they lost their validity on the subsequent two occasions, but gained it again 1 month after the birth; the optimal cut-off points varied accordingly. This study suggests that the optimal cut-off point for a questionnaire should be validated against an externally determined clinical diagnosis whenever the instrument is used repeatedly on the same population.

T. Kitamura [1], S. Shima [2], M. Sugawara [3], M. A. Toda [4]

[1] Department of Sociocultural Environmental Research, NIMH, Ichikawa, [2] Tokyo Keizai University, Tokyo, [3] North Shore College, Atsugi, [4] Hokkaido University of Education, Sapporo, Japan

Key words: questionnaire; validity; depression; neurosis; pregnancy; puerperium; rating scale

T. Kitamura, Department of Sociocultural Environmental Research, NIMH, 1-7-3 Konodai, Ichikawa, Chiba 272, Japan

Accepted for publication August 6, 1994

Self-rating questionnaires have frequently been used in both epidemiological and clinical studies when the samples are of a relatively large size or of a nature that does not allow direct access to the subjects.

The validity of self-rating questionnaires is measured in terms of sensitivity (the proportion of cases correctly identified by positive scores of the questionnaire to true cases), specificity (the proportion of non-cases (normal subjects) correctly identified by negative scores of the questionnaire to true non-cases), positive predictive value (the proportion of true cases to those with positive scores of the questionnaire), negative predictive value (the proportion of true non-cases to those with negative scores) and overall diagnostic rate (the proportion of cases and non-cases correctly identified by positive and negative scores respectively to the whole sample). The validity of self-rating questionnaires is usually expressed in terms of these 5 measures. However, the power of the last 3 measures varies, depending on the base rate (prevalence) of cases, whereas sensitivity and specificity do not. Therefore, these two measures are usually used to indicate the screening power of a questionnaire, whereas for the purposes of epidemiological studies, positive and negative predictive values are more useful in estimating the prevalence of cases. An optimal cut-off point is specified for a questionnaire, taking into account the balance of the 5 validity measures. The external criterion for validation is usually a clinical judgement, simultaneously made by experienced clinicians or researchers who are blind to the questionnaire score. However, it is usual to examine the validity of a questionnaire for use on a single occasion, whereas its reliability is often measured by comparing the scores recorded on two different occasions. Although the validity of repeatedly administered questionnaires cannot be endorsed by a study of their use on single occasions, this has rarely been examined.

We (1, 2) have already found that, among pregnant women, the total score of the Japanese version of the GHQ satisfactorily discriminates psychiatric cases and non-cases at the first trimester and that the validity figures of the GHQ at the third trimes-

ter became poor. This article is an extended report of the validity of the repeated use of two well-known self-rating psychiatric questionnaires – the General Health Questionnaire (GHQ) (3) and Zung's Self-rating Depression Scale (SDS) (4) – in women passing through both the antenatal and postnatal periods.

Material and methods

The sample for this present study consisted of 120 women who volunteered to participate in a follow-up study of their mental health during both the antenatal and postnatal periods (1, 2). At entry, they were all attending the Department of Obstetrics of a general hospital in Kawasaki, Japan; women at over 12 weeks gestation were excluded, but no other pregnant women were excluded. They were aged 17 to 42 years, with a mean (SD) of 28 (5) years. All were married. The pregnancy was their first for 42 (35%) women; 66 (55%) women were expecting their first baby. Social class was not examined, because of the lack of an agreed definition for it in Japan.

The women were examined four times: (a) in early pregnancy (when the fetal heart beat was first confirmed by echocardiography); (b) in late pregnancy (approximately 34 weeks gestation); (c) 5 days postnatally; and (d) 1 month postnatally. At each examination, the 30-item GHQ and SDS were administered.

The GHQ aims to identify current nonpsychotic mental disorder in general. The original 60-item version has been shortened to 30-, 28-, 20- and 12-item questionnaires, but the 30-item version has been selected for use by many researchers. Each item is rated as either 0 or 1; the total score of the 30-item GHQ ranges between 0 and 30, with higher scores meaning a higher probability of caseness. The SDS was developed as an instrument to assess the severity of depression but is often used as a screening instrument. Unlike the original SDS, scores of 0 to 3 were assigned to the 4 responses of each SDS item, so that the range of the total SDS score became 0–60, rather than 20–80, as in the original version.

The subjects were then interviewed by one of the two psychiatrists (T.K. and S.S.) blind to the results of the two questionnaires. These interviews were conducted using the Schedule for Affective Disorders & Schizophrenia (SADS) (5) or its change version (6) in the second and later interviews, and psychiatric diagnosis was established according to the Research Diagnostic Criteria (RDC) (7). The reliability of the RDC and SADS had been confirmed by both case vignettes (8) and interrater designs (9), prior to the present study; their interrater reliability coefficients expressed by Cohen's (10) kappa for RDC major depressive disorder were 1.00 and 0.84 in the case vignette and interrater designs, respectively.

For the validation study of the GHQ, the women who were assigned any of the RDC diagnoses were designated as cases at each time-point, and those without any RDC diagnosis were designated non-cases. The discriminant power of the total GHQ score was assessed against the caseness, and mean scores of the GHQ compared between cases and non-cases at each time-point. Similarly, the SDS was tested against RDC depressive disorders (major and minor depressive disorders and other mental disorder (OMD) with dysphoric mood) as the external validator. For these calculations the SPSS-X program (11) was used.

Table 1. Validity of GHQ for RDC cases

GHQ cut-off points	First trimester		Third trimester		Postnatal day 5		Postnatal month 1	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
0/1	100.0	3.3	92.3	12.9	100.0	8.3	100.0	4.8
1/2	100.0	11.1	76.9	29.4	100.0	31.9	93.8	27.7
2/3	94.4	31.1	61.5	49.4	77.8	43.1	81.3	39.8
3/4	88.9	38.9	38.5	56.5	61.1	54.2	81.3	54.2
4/5	88.9	47.8	38.5	64.7	44.4	63.9	68.8	61.4
5/6	88.9	54.4	38.5	72.9	27.8	68.1	50.0	68.7
6/7	83.3	64.4	38.5	77.6	27.8	73.6	50.0	75.9
7/8	83.3	71.1	38.5	82.4	27.8	79.2	50.0	84.3
8/9	77.8	75.6	30.8	83.5	22.2	80.6	37.5	85.5
9/10	77.8	78.9	23.1	87.1	11.1	81.9	37.5	88.0
10/11	66.7	78.9	23.1	88.2	11.1	86.1	37.5	89.2
11/12	66.1	83.3	7.7	91.8	11.1	87.5	37.5	89.2
12/13	44.4	87.8	7.7	95.3	11.1	88.8	37.5	89.2
13/14	27.8	92.2	7.7	95.3	11.1	90.2	37.5	90.4
14/15	27.8	93.3	7.7	95.3	11.1	95.8	25.0	94.0

Results

Usable questionnaires were returned for the GHQ from 108, 98, 90, and 99 women at the first and third trimesters, and 5 days and 1 month postnatally, respectively. The corresponding figures for the SDS were 111, 102, 91, and 101. The number of women whom we did not manage to interview was 15, 17 and 15 at the third trimester, 5 days, and 1 month postnatally; 2 women experienced neonatal death, 3 experienced miscarriage, and the remaining women either moved out of the area or did not attend. Therefore, the questionnaire response rates among those women interviewed were high.

In the first trimester, the RDC diagnoses were; major depressive disorder, 9; minor depressive disorder, 3; OMD with dysphoric mood, 1; phobic disorder, 2; obsessive-compulsive disorder, 3; panic disorder, 1; labile personality, 1; OMD without dysphoric mood, 3; and schizotypal features, 1. In the third trimester, the RDC diagnoses were: major depressive disorder; 5; minor depressive disorder; 5; OMD with dysphoric mood; 1; phobic disorder; 2; obsessive-compulsive disorder; 5; and labile personality, 1. On day 5 postnatally, the RDC diagnoses were; minor depressive disorder, 6; OMD with dysphoric mood, 6; phobic disorder, 2; obsessive compulsive disorder, 4; labile personality, 1; and OMD without dysphoric mood, 2. At 1 month postnatally, the RDC diagnoses were; major depressive disorder, 3; minor depressive disorder, 4; OMD with dysphoric mood, 3; phobic disorder, 2; obsessive-compulsive disorder, 7; labile personality, 1; and OMD without dysphoric mood, 1. The sum of the number of cases described above exceeds the number of cases indicated in Table 2 because of the multiple diagnosis policy of the RDC. OMD is a disorder that shows some mental symptoms but does not meet any of RDC diagnoses.

Since the purpose of the GHQ is to identify cases with nonorganic nonpsychotic minor mental disorder in general, whereas the SDS identifies cases of depressive disorder (3), the external criteria against which the GHQ and SDS were validated were, for the GHQ, all RDC diagnoses, and for the SDS, depressive disorders. In this study, depressive disorders included major and minor depressive disorders and OMD with dysphoric mood. The validity of the GHQ at the first trimester has been reported previously elsewhere (1).

At the first trimester, we set an optimal cut-off point of the total GHQ score at 7/8, taking into account all 5 validation measures (1). The cut-off point of 7/8, however, was unable to maintain the scale's sensitivity over the subsequent three examination time-points, though the specificity remained stable throughout (Table 1). Sensitivity could be retained by reducing the cut-off point, but this substantially invalidated the scale's specificity. However, the sensitivity of the total GHQ score seemed to increase again at 1 month postnatally.

The differences of mean GHQ scores between cases and non-cases were most significant at the first trimester ($t = 4.58$, $P = 0.001$), lost significance at the third trimester and day 5 postnatally, but gained significance again at 1 month postnatally ($t = 2.43$, $P = 0.025$) (Table 2).

Inspection of the sensitivity and specificity of the total SDS score indicates that its optimal cut-off point at the first trimester may be set at 22/23 (Table 3). Like the total GHQ score, the total SDS score lost its sensitivity as the follow-up continued, but retained it slightly at 1 month postnatally. The specificity of the SDS was more or less stable over the 4 time-points.

Differences of the mean SDS scores between cases and non-cases were significant at the two antenatal time-points but lost significance postnatally (Table 4).

The above findings were virtually the same when cases with OMD were redesignated as non-cases (tables not shown here, but available on request to the first author).

Discussion

The salient findings of this study are that in the first trimester both the GHQ and SDS identified cases of minor mental disorder and depressive disorders respectively, that they lost their validity on the subsequent two occasions, but gained it again 1 month postnatally. The optimal cut-off points varied accordingly.

Previous reports indicate that the GHQ and SDS (12–14) are powerful means of identifying psychiatric cases or measuring severity. However, most previous studies compared the results of each questionnaire against a psychiatric diagnosis made con-

Table 2. GHQ scores of the RDC cases and controls

	First trimester	Third trimester	Postnatal day 1	Postnatal month 1
Cases				
n	18	13	18	16
mean	12.3	5.8	5.9	9.3
SD	5.7	6.5	5.1	7.3
Non-cases				
n	90	85	72	83
mean	6.2	4.2	4.7	4.7
SD	5.1	4.2	4.5	4.4
t	4.58	0.87	1.02	2.43
P	<0.001	NS	NS	<0.05

t: two-tailed Student's *t*-test.

Table 3. Validity of the SDS for RDC depressive disorders

SDS cut-off points	First trimester		Third trimester		Postnatal day 5		Postnatal month 1	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
9/10	100.0	3.0	100.0	3.3	100.0	7.5	88.9	5.4
10/11	100.0	4.0	100.0	4.3	100.0	11.3	88.9	8.7
11/12	100.0	6.0	100.0	6.5	100.0	16.3	88.9	12.0
12/13	100.0	9.0	100.0	6.5	90.9	18.8	88.9	22.8
13/14	100.0	14.0	90.0	10.9	81.8	31.3	77.8	29.3
14/15	100.0	17.0	90.0	15.2	72.7	40.0	77.8	34.8
15/16	100.0	23.0	90.0	27.2	72.7	51.3	77.8	47.8
16/17	100.0	29.0	90.0	39.1	63.6	55.0	77.8	56.5
17/18	100.0	37.0	90.0	47.8	54.5	61.3	77.8	63.0
18/19	100.0	43.0	80.0	53.3	36.4	70.0	77.8	72.8
19/20	100.0	51.0	80.0	60.9	27.3	76.3	55.6	80.4
20/21	100.0	57.0	80.0	64.1	27.3	77.5	55.6	83.7
21/22	90.9	62.0	80.0	67.4	27.3	80.0	55.6	87.0
22/23	90.9	70.0	70.0	76.1	27.3	85.0	44.9	88.0
23/24	72.7	74.0	50.0	82.6	27.3	90.0	33.3	92.4
24/25	72.7	78.0	50.0	89.1	18.2	92.5	22.2	94.6
25/26	72.7	85.0	40.0	91.3	18.2	93.8	11.1	95.7
26/27	63.6	88.0	40.0	92.4	18.2	95.0	11.1	96.7
27/28	63.6	91.0	20.0	92.4	9.1	96.3	11.1	96.7
28/29	54.5	91.0	10.0	95.7	9.1	97.5	11.1	96.7
29/30	45.5	93.0	10.0	96.7	9.1	97.5	11.1	98.9

currently by clinicians blind to the questionnaire scores. Our results strongly suggest that the validity of these two well-known questionnaires, and possibly of any questionnaire, varies on different occasions.

The fact that the validity of the questionnaires is reduced on the subsequent two occasions may be interpreted as the effect of social desirability, enhanced by repeated exposure to the same stimuli. Thus, using the same sample, we have demonstrated that the reduced difference in the mean GHQ scores between cases and non-cases from the first to second occasions was not due to an increase in the GHQ scores of the non-cases but due to a decrease in the GHQ scores of the cases (2). Those women with an initially high GHQ score may have become aware of the purpose of GHQ questions when faced by them on the second occasion.

Any effect of social desirability, however, could not explain the fact that 1 month after childbirth, the validity of the questionnaires was as high as that in the first trimester. A possible explanation may be that the nature of cases in the first trimester and 1 month postnatally is different from that in the third trimester and 5 days postnatally. Thus, of 13 depressive cases in the first trimester, 9 were major depressive disorder; of 10 depressive cases at 1 month postnatally, 3 were major depressive disorder, whereas only one had major depressive disorder on day 5 postnatally. Though there were 5 cases with major depressive disorder in the third trimester, the validity of the questionnaires was reduced markedly. Therefore, the different nature of diagnoses on different occasions cannot be the sole explanation.

It is usual practice in Japan for pregnant women to attend an obstetric clinic regularly (often once a month) during the antenatal period, but after childbirth, they attend there only at 1 month and rarely thereafter. It may be speculated that their attitude to a questionnaire is different between the antenatal and postnatal periods, and that the effect of social desirability is more prominent during the antenatal period, possibly because they more strongly depend on the care delivered by hospital staff.

Because we can neither predict how much social desirability works on the response of the subjects nor forsee the symptomatic and diagnostic structure

Table 4. SDS scores of women with RDC depressive disorders and controls

	First trimester	Third trimester	Postnatal day 1	Postnatal month 1
Cases				
n	11	10	11	9
mean	28.4	24.0	19.0	21.2
SD	4.9	5.8	6.0	8.8
Non-cases				
n	100	92	80	92
mean	19.9	18.8	16.6	16.3
SD	5.9	5.2	5.4	4.8
t	4.61	3.02	1.38	1.65
P	<0.001	<0.01	NS	NS

t: two-tailed Student's *t*-test.

of cases in a given population, there seems no way to set an optimal cut-off point for a psychiatric questionnaire, other than to validate it against an externally determined clinical diagnosis, even when it is used on the same population.

Acknowledgements

We wish to thank S. Hayashi and K. Sakakura, Department of Gynaecology and Obstetrics and T. Shikano, Department of Psychiatry, Kawasaki Municipal Hospital, Kawasaki, Japan, for their help. The present research was partly supported by a grant from the Toyota Foundation (86-II-081).

References

1. KITAMURA T, SUGAWARA M, AOKI M, SHIMA S. Validity of the Japanese version of the GHQ among antenatal clinic attendants. Psychol Med 1989: 19: 507–511.
2. KITAMURA T, TODA MA, SHIMA S, SUGAWARA M. Validity of the repeated GHQ among pregnant women: a study in a Japanese general hospital. Int J Psychiatry Med (in press).
3. GOLDBERG DP. The detection of psychiatric illness by questionnaire: a technique for the identification and assessment of non-psychotic psychiatric illness. Maudsley Monograph No. 21. Oxford: Oxford University Press, 1972.
4. ZUNG WK, DURHAM NC. A self-rating scale. Arch Gen Psychiatry 1965: 12: 63–70.
5. SPITZER RL, ENDICOTT J. Schedule for Affective Disorders and Schizophrenia (SADS). 3rd edn. New York: Biometric Research, New York State Psychiatric Institute, 1978.
6. SPITZER RL, ENDICOTT J. Schedule for Affective Disorders and Schizophrenia: change version (SADS-C). 3rd edn. New York: Biometric Research, New York State Psychiatric Institute, 1978.
7. SPITZER RL, ENDICOTT J, ROBINS E. Research Diagnostic Criteria (RDC) for a selected group of functional disorders. New York: Biometric Research, New York State Psychiatric Institute, 1978.
8. KITAMURA T, SHIMA S, SAKIO E, KATO M. Application of Research Diagnostic Criteria and International Classification of Diseases to case vignettes. J Clin Psychiatry 1986: 47: 78–80.
9. KITAMURA T, SHIMA S. [Interrater reliability of the Schedule for Affective Disorders and Schizophrenia (SADS) and Research Diagnostic Criteria (RDC).] Seishin Igaku (Clin Psychiatry) 1986: 28: 41–45 (in Japanese).
10. COHEN J. A coefficient of agreement for moninal scales. Educ Psychol Meas 1960: 20: 37–46.
11. SPSS. SPSS user's guide. 2nd edn. Chicago: SPSS, 1986.
12. CARROLL BJ, FIELDING JM, BLASHKI TG. Depression rating scales. Arch Gen Psychiatry 1973: 28: 361–366.
13. DAVIES B, BURROWS G, POYNTON C. A comparative study of four depression rating scales. Aust NZ J Psychiatry 1975: 9: 21–23.
14. SNYDER S, PITTS WM. Comparison of self-rated and observerrated scales in DSM-III borderline personality disorder. Can J Psychiatry 1986: 31: 708–713.