The translation into Arabic and revalidation of a fatigue questionnaire

S. McIlvenny,[1] M.H. Ahmed,[2] E. Dunn,[1] H. Swadi[3] and M. Balshie[1]

ترجمة استبيان حول التعب إلى اللغة العربية ثم إعادة التحقُّق من مصداقيته

شيرلي ماكلفني ومحمد الصادق حاج أحمد وإيرل فنسنت دان وحارث سوادي ومحمد جميل بلشة

خلاصة: كان هدف هذه الدراسة ترجمة استبيان حول التعب أُعِدَّ في بريطانيا ليُستعمل في الدراسات الوبائية وفي الأوضاع المجتمعية، إلى اللغة العربية. وكان المقصود أن يمكن استعمال الاستبيان المترجم في أي وضع تكون اللغة العربية فيه هي اللغة الأولى التي يتحدث بها المرضى. وتصف هذه المقالة كيفية ترجمة الاستبيان وطريقة إعادة التحقق من مصداقيته. ولقد ظهر أن الترجمة العربية تتَّسم بالموثوقية والمصداقية لدى استعمالها في الإمارات العربية المتحدة.

ABSTRACT The aim of this study was to translate a fatigue questionnaire, which had been developed in England for use in epidemiological studies and in community settings, into Arabic. It was intended that the translated questionnaire could be used in any setting where Arabic is the first language of the patient. The process of translating the questionnaire and the revalidation method are described. The Arabic translation was shown to be both reliable and valid in the United Arab Emirates setting.

Traduction en arabe et revalidation d'un questionnaire relatif à la fatigue

RESUME Le but de cette étude était de traduire en arabe un questionnaire relatif à la fatigue qui avait été mis au point en Angleterre pour être utilisé dans des études épidémiologiques et dans des établissements communautaires. On projetait d'utiliser le questionnaire traduit dans tout établissement où l'arabe est la langue principale utilisée par le patient. Le processus de traduction du questionnaire et la méthode de revalidation sont décrits. La traduction en arabe s'est avérée être fiable et valable dans le contexte des Emirats arabes unis.

[1]Department of Family Medicine; [2]Department of Community Medicine; [3]Department of Psychiatry, Faculty of Medicine and Health Sciences, United Arab Emirates University, Al-Ain, United Arab Emirates.
Received: 27/01/99; accepted: 18/03/99

Introduction

Fatigue is the seventh most common complaint seen at primary health care centres and the usual method of studying this phenomenon is by a questionnaire regarding the symptoms associated with fatigue [*1*]. In order to study fatigue in an Arabic-speaking population, an English questionnaire was translated into Arabic and revalidated. The aim of this study was to describe the method of revalidation and the results obtained.

The questionnaire

The fatigue scale, developed by Chalder et al., was chosen to study the prevalence of fatigue and permission was obtained from the authors [*2*]. The questionnaire contained 11 questions, 7 physical and 4 mental fatigue items. Patients were asked if they had experienced each symptom in the last 2 weeks and the questionnaire was scored using a bimodal response system (0,0,1,1). The fatigue score was calculated by adding together all the item scores (maximum score 11) and a total of four or more constituted a case of fatigue. The questionnaire is shown in Figure 1.

Fatigue questionnaire

Physical symptoms

1. Do you have problems with tiredness ?
2. Do you need to rest more ?
3. Do you feel sleepy or drowsy ?
4. Do you have problems starting things ?
5. Are you lacking in energy ?
6. Do you have less strength in your muscles?
7. Do you feel weak ?

Mental symptoms

8. Do you have difficulty concentrating ?
9. Do you have problems thinking clearly ?
10. Do you make slips of the tongue when speaking ?
11. Do you have problems with your memory ?

A bimodal response system was used as follows: less than usual (0); same as usual (0); more than usual (1); much more than usual (1).

Figure 1 **Fatigue scale developed by Trudie Chalder [*2*]**

Method

The questionnaire was translated into Arabic separately by two translators. These two versions were combined and revised and then back-translated into English by a third translator. The bilingual administrators used were native Arabic speakers and resident in the United Arab Emirates (UAE). They were selected from different Arabic-speaking countries and differing levels of education. The translation was refined after the back translation until agreement was obtained among the administrators.

The translation was then piloted for comprehension and ease of administration on a population of consecutive UAE national primary health care (PHC) patients attending the Family Medicine Clinic, Tawam Hospital, for any reason. As UAE patients are not familiar with completing questionnaires and 15% of the population is illiterate, it was decided to have all questionnaires administered by nurse-interpreters working in the clinic. The translation was again refined and a final version agreed. The resulting Arabic fatigue scale is shown in Figure 2.

الأعراض الجسمانية

خلال الأسبوعين الماضيين:	أقل من المعتاد	عادي	أكثر من المعتاد	أكثر كثيراً من المعتاد
1. هل عانيت من متاعب بسبب الإرهاق؟	_____	_____	_____	_____
2. هل رغبت في أخذ راحة أكثر؟	_____	_____	_____	_____
3. هل شعرت بالنعاس أو الخمول؟	_____	_____	_____	_____
4. هل شعرت بصعوبة في ابتداء شيء ما؟	_____	_____	_____	_____
5. هل شعرت بنقص في طاقتك؟	_____	_____	_____	_____
6. هل عانيت من ضعف بعضلاتك؟	_____	_____	_____	_____
7. هل حدث أن شعرت بضعف عام؟	_____	_____	_____	_____

الأعراض العقلية:

8. هل لديك صعوبة في التركيز؟	_____	_____	_____	_____
9. هل لديك صعوبة في التفكير بوضوح؟	_____	_____	_____	_____
10. هل تحدث لديك زلة لسان عند الكلام؟	_____	_____	_____	_____
11. ما هي حالة ذاكرتك؟	_____	_____	_____	_____

Figure 2 **Arabic fatigue scale**

Determining the cut-off point on the fatigue scale

The Arabic version was compared with the same gold standard that was used to validate the English version. This was the relevant item on fatigue in the Revised Clinical Interview Schedule (CIS-R) [*3*]. The item required a positive answer to one of two questions ("Have you noticed that you've been getting tired recently?" and "Have you felt that you've been lacking in energy?"). If one response to this item was yes, it was followed by four supplementary questions, as shown in Figure 3. A positive response to each question scored one. A total score of two or more (maximum score four) was considered a fatigue case. The gold standard was translated into Arabic, then both the questionnaire and the gold standard were administered to a sample of PHC patients. These patients were randomized, using random number tables, to receive either the questionnaire or the gold standard first. The cut-off point on the fatigue scale was determined using relative operating characteristic (ROC) curve analysis; the same method as the English questionnaire [*4*].

Validity

A group of experts, consisting of five family physicians, examined the English questionnaire for content and construct validity. The expert panel examined the structure of the questionnaire and decided that it covered all aspects of fatigue, both physical and mental. The marking of the scale was examined and the weighting of the scores discussed.

Fatigue item

Have you noticed that you've been getting tired recently ? yes/no

Have you felt that you've been lacking in energy ? yes/no

If the answer to either question is yes, answer the next four questions.

1. On how many days have you felt tired or lacking in energy during the past week? (Score 1 for greater than or equal to 4 days)
2. Have you felt tired/lacking in energy for more than 3 hours on any day in the past week? (score 1 for greater than 3 hours)
3. Have you felt so tired that you've had to push yourself to get things done during the past week? (score 1 for yes)
4. Have you felt tired or lacking in energy when doing things you enjoy during the last week? (score 1 for yes)

This gives a total fatigue score from 0–4. Scores of 2 or above are regarded as fatigued.

Figure 3 **English version of the gold standard [*3*]**

validity. Each item was discussed and the questionnaire was assessed for its relevance to the study setting. Face validity was qualitatively assessed by a lay panel of nine bilingual UAE medical students who were also members of the target population of patients attending the clinic. The group discussed the various Arabic words describing fatigue and the symptoms patients would normally associate with fatigue.

Criterion validity was formally tested by administering the English and Arabic versions to a population of bilingual UAE medical students and comparing the results. Students were randomized to receive either the English or Arabic version first and completed the questionnaire themselves. The total fatigue scores and scores for physical and mental fatigue were then compared.

Internal reliability

The internal reliability of the questionnaire was tested using Cronbach alpha. This test was carried out on the sample of patients used to calculate the cut-off point. Cronbach alpha was calculated for all items and after taking out each item one at a time. The result was compared with the English version. The patient sample was divided into males and females and for each gender group Cronbach alpha was again calculated for each item removed from the scale.

A new population of medical students was used to examine the test–retest reliability of the Arabic questionnaire with 1 hour between each application. Inter-observer bias was tested using a sample of consecutive PHC patients and six research assistants, two male and four female. The participants were administered the questionnaire twice, first by one research assistant then by a second, 1 hour later.

Results

All items in the fatigue scale were considered easy to translate from English to Arabic and in most cases direct translations of the English phrases were obtained.

Pilot study

Thirteen (13) male and 6 female PHC patients were given the fatigue scale by one male and one female research assistant. One male patient refused consent. The ages of the participants ranged from 19 years to 72 years (mean 45.2 years, median 38.5 years). Scores ranged from 1 to 9 for males and 4 to 10 for females. The mean score for males was 3.50 and for females 6.83. The questionnaire was well received by the patients and the research assistants who were interviewed after the pilot study. Minor adjustments were made to the questionnaire to improve clarity but no major changes were judged necessary.

Determining the cut-off point on the scale

Sixty subjects (60), 26 males and 34 females, received both the Arabic questionnaire and the gold standard. ROC curve analysis showed the best cut-off point to be between 4 and 5, where a score greater than or equal to five is said to indicate fatigue. This is one point higher than for the English questionnaire, where the cut-off point is between 3 and 4. This was believed to be due to cultural factors rather than language issues. The results are shown in Table 1.

Criterion validity

The English and Arabic versions were compared using 46 students, 16 males and 30 females, who completed both the English and Arabic questionnaires. Twenty-three (23) received the Arabic version first and 23 received the English first. A paired *t*-test

Table 1 **ROC curve analysis comparing fatigue score with gold standard**

Score	Sensitivity (%)	95% CI for sensitivity	Specificity (%)	95% CI for specificity
≥0	100.0	100.0–100.0	0.0	0.0–0.0
>0	100.0	100.0–100.0	25.0	11.5–43.4
>1	92.6	75.7–98.9	34.4	18.6–53.2
>2	81.5	61.9–93.6	50.0	31.9–68.1
>3	77.8	57.7–91.3	65.6	46.8–81.4
>4[a]	59.3	38.8–77.6	87.5	71.0–96.4
>5	40.7	22.4–61.2	87.5	71.0–96.4
>6	37.0	19.4–57.6	93.8	79.2–99.1
>7	14.8	4.3–33.7	100.0	100.0–100.0
>8	7.4	1.1–24.3	100.0	100.0–100.0
>9	3.7	0.6–19.0	100.0	100.0–100.0
>10	0.0	0.0–0.0	100.0	100.0–100.0

[a] *cut-off point*
ROC = relative operating characteristic
CI = confidence intervals

was carried out comparing responses but there was no significant difference between the Arabic and English versions (P = 0.694.) The P-values, mean difference, confidence intervals and Kappa values are shown in Table 2. The fatigue scores were converted into yes/no results for fatigue based on the calculated cut-off point on the Arabic scale. Cross tabulations were made for each test and the Kappa value calculated at 0.445. The results for physical fatigue and mental fatigue were compared. Again paired t-test analysis showed no significant difference between the Arabic and English versions.

Item precision

The precision for each item was calculated and a comparison made between males and females. The range of precision was between 0.75 and 0.94 for males and 0.73 and 0.97 for females and the results are shown in Table 3.

Internal consistency of the questionnaire

Cronbach alpha, calculated for all items, was 0.7486. This was repeated after taking out each item one at a time and the results are shown in Table 4. The results ranged from 0.6930 to 0.7491. This was lower than the English fatigue scale which ranged from 0.88 to 0.90. This was repeated on males and females separately as a comparison. The results were 0.7668 for males (range 0.7075–0.7770) and 0.7413 for females (range 0.6897–0.7537).

Test–retest reliability

Twenty-eight students (28), 8 males and 20 females, completed the test–retest procedure on the Arabic questionnaire. There was a significant difference between fa-

Table 2 **Results for paired *t*-tests and Kappa values**

Test	No. pairs	Mean	Mean difference	95% CI for mean difference	*P*-value	Kappa value		
						Male	Female	Total
Arabic/English fatigue score	46	2.76 2.86	0.108	–0.662 to 0.445	0.694	0.448	0.426	0.445
Arabic/English physical fatigue score	46	1.96 2.17	0.217	–0.673 to 0.238	0.341			
Arabic/English mental fatigue score	46	0.826 0.696	0.130	–0.109 to 0.370	0.278			
Test/retest	28	3.39/2.57	0.821	0.179 to 1.464	0.014	1.0	0.494	0.588
Inter-rater reliability	25	5.32 3.92	1.400	0.574 to 2.226	0.002	1.0	0.333	0.525

CI = confidence intervals

tigue scores on the paired *t*–test (P = 0.014). Scores were again converted into yes/no results for fatigue based on the calculated cut-off point. In cross tabulations the Kappa value was 0.588.

Inter-rater reliability

Twenty-five (25) general practitioner patients, 7 males and 18 females, were given the questionnaire by two research assistants. There was a significant difference between fatigue scores on the inter-rater reliability (P = 0.002). Scores were converted into yes/no results for fatigue and cross tabulations showed the Kappa value was 0.525.

Discussion

The questionnaire was translated without difficulty into Arabic and required only a small degree of interpretation during its administration on the part of the research assistants. Even though bilingual research assistants were used, the questionnaire was translated into Arabic to reduce observer interpretation during administration. This was especially important as the setting was a very different culture from the one in which the original questionnaire had been developed.

Various techniques are available to the interpreter when translating a questionnaire and these are usually divided into two categories — direct and indirect translation. In this case, many nouns and verbs were translated directly from English to Arabic and there were no words which did not have an Arabic equivalent. In the indirect translation, transposition and modulation were used. Transposition occurs when one grammatical part of speech is substituted for another without changing the meaning. As Arabic is grammatically very different from

Table 3 **Precision for each item of the scale in the criterion validity and reliability tests**

Item	Criterion validity		Test–retest reliability		Inter-rater reliability	
	Male	Female	Male	Female	Male	Female
1	0.75	0.8	1.0	0.75	1.0	0.89
2	0.75	0.83	1.0	0.6	1.0	0.83
3	0.81	0.73	0.87	0.8	1.0	0.78
4	0.87	0.83	1.0	0.8	1.0	0.78
5	0.75	0.73	0.75	0.85	1.0	0.38
6	0.87	0.83	1.0	0.9	1.0	0.83
7	0.81	0.76	0.87	0.8	1.0	0.72
8	0.87	0.83	0.87	0.7	1.0	0.67
9	0.94	0.87	0.87	0.8	1.0	0.83
10	0.87	0.87	1.0	0.9	1.0	1.0
11	0.87	0.97	0.87	0.85	1.0	0.67

Table 4 **Cronbach alpha value for each item: a comparison of males and females**

Item	Males	Females
1	0.7526	0.6897
2	0.7620	0.7326
3	0.7526	0.7198
4	0.7380	0.7498
5	0.7075	0.6923
6	0.7381	0.6996
7	0.7281	0.6989
8	0.7637	0.7461
9	0.7596	0.7234
10	0.7489	0.7537
11	0.7770	0.7240

English, the translation involved transposition of every item. Modulation occurs when a fixed expression or cliché is used instead of direct translation to convey the same meaning. This technique was used for several items; for example, the English expression "slip of the tongue" was replaced by one word in Arabic which conveys the same meaning.

Hui and Triandis postulated four levels of equivalence in which a questionnaire must be comparable to the original version [5]. Each level is a prerequisite of the next level as follows:

- conceptual/functional equivalence
- construct equivalence
- item equivalence
- scalar equivalence.

In conceptual equivalence, the responses must reflect the same concept as the original authors intended. In the fatigue scale, conceptual equivalence was agreed by the translators, the students who took part as subjects and the nurse interpreters.

Construct equivalence refers to the definition of the construct (medical condition) to be measured. In other words, does the construct agree with results from other questionnaires? Does it predict high scores

in fatigued patients and low scores in non-fatigued patients in the same way as the English version? Construct equivalence was supported by the comparison with the gold standard. This showed moderate agreement between the two scales with a cut-off point that was only one point higher than in the English version.

In item equivalence, items can be equivalent in meaning and concept. "Culture-linked" items may be equivalent in meaning but do not measure the same concept in different settings. In the translation of the fatigue scale, 10 items were equivalent in meaning and concept. However question 5, "Are you lacking in energy?", posed a problem. In general usage, the word "energy" in Arabic is related to "fitness" as in "I am feeling fit" rather than "I am full of energy". The same word is used to denote tiredness as in "I am not feeling fit". However, there is a specific word for "energy" in Arabic, used more for the concept of scientific energy, and this was employed in the translation. The term "energy" and its conceptual meaning in Arabic would suggest that the item "Are you lacking in energy?" is a culture-linked item.

In scalar equivalence, the responses to each item should have the same rank order and interval score properties as the source questionnaire. For instance, if the item has a Likert-type scale, then the difference between two points on the scale should be equidistant on the target scale and the source scale. With regard to the fatigue scale, the paired *t*-test results for criterion validity gave support to the hypothesis that the two scales, English and Arabic, had scalar equivalence.

The cut-off point indicated on the basis of the ROC curve analysis was between 4 and 5, one point higher than the English fatigue scale. The cut-off point on a scale should also be a balance between sensitivity and specificity and this balance depends on the use the instrument is designed for. In clinical practice, it is desirable to have a high specificity which minimizes false positives in order to exclude patients whose symptoms are so mild that they do not require treatment. This is especially true for a symptom like fatigue which is viewed as a continuum, occurring in normal people and ill patients alike. The cut-off point in this case gave a high specificity of 87.5% (decreasing the false positives) at the expense of a lower sensitivity of 59.3%.

The point must be raised that the gold standard used is not necessarily a valid measure to compare the questionnaire with. No questionnaire or laboratory test has been developed which is an absolute measure of the perception of fatigue. The gold standard which was used by the original authors had not been validated or tested for reliability itself, a point raised earlier by the authors. Furthermore, the Arabic translation of the gold standard was not validated or shown to be a reliable translation of the original version. However, the general consensus of the medical experts and the lay panel was that the gold standard was an adequate measure with which to compare the scale and thus the Arabic version of the gold standard also had face validity. When scores were converted into yes/no for fatigue, there was general agreement between the gold standard and the fatigue scale, with 22 out of 26 males and 23 out of 34 females agreeing on fatigue status.

The content, construct and face validity of the Arabic fatigue scale were judged to be adequate by the expert and lay panels. Criterion validity was assessed using a comparison of the English and Arabic questionnaires. There was no significant difference between the two scales for the total fatigue score, physical or mental fatigue scores, and the Kappa value showed moder-

ate agreement between the two versions in detecting fatigued patients. In the item analysis, females tended to have fewer symptoms in the Arabic version than the English version. No such finding occurred in the male subjects. Is the Arabic scale minimizing the symptoms of females? The consistency of the male responses indicates that this was probably not a difficulty with the language used in the questionnaire and suggests that there is a cultural difference in the way females respond to the questions.

Reliability of the questionnaire

The internal consistency of the Arabic fatigue scale was comparable to the English version and did not indicate that any item could be eliminated from the scale. The Cronbach alpha was consistent for each item, although the results were lower than for the English version.

The test retest analysis showed a significant *P*-value in the paired *t*-test analysis. However, the confidence interval was relatively small and the Kappa value supported general agreement after 1 hour. Although the questionnaire asked about symptoms over the preceding 2 weeks, responses may have been affected by how the subject was feeling at the time of the test. One hour was agreed as the maximum time lapse allowable between tests, although the patient might still feel less fatigued even after this short interval of resting or intake of food and fluids. The Kappa value for males in the test retest analysis showed total agreement, thus supporting the concept that the questionnaire may be generalized over different occasions. In females, the Kappa value decreased to 0.494 and four females who were fatigued at the first administration had recovered by the second test. Also in the item analysis, females changing from a positive to a negative response after 1 hour was noticeable in nine items.

The inter-rater reliability test, in contrast, was conducted on general practitioner patients attending a routine, mid-week clinic. Again the *P*-value was significant and the confidence interval had increased, indicating that some variance may have been due to the research assistants. However, the Kappa value showed moderate agreement between observers in detecting fatigue cases. It is possible that the observers did show some bias in their interpretation of the responses, but these tests were carried out immediately after their initial training. The observers improved their skills with time and it is anticipated that the inter-rater effect diminished with practice. Nevertheless, it is more likely that what was being observed was a test–retest effect as in the previous test.

In the item analysis, the precision in males for all items was 1.0, indicating good inter-rater reliability. The same phenomenon occurred, as in the test–retest, where four females who were fatigued in the first administration had recovered by the second administration. The item analysis showed that in eight items at least two females changed from a positive to a negative response with the second researcher. It is interesting to observe that the two items, "Are you lacking in energy?" and "Do you have difficulty concentrating?", were particularly susceptible to change in females, in both test–retest and inter-rater reliability.

Gender differences in responses to questionnaires have been studied in other cultures but no such studies have been carried out in Arabic populations. Three reasons have been postulated for these differences: gender stereotypes, external influences and physiological factors [*6*]. Gender stereotypes may be affected by cultural beliefs, in that patients behave in a certain way in keeping with cultural expectations. In the Arab culture, there is a com-

monly held belief that women change their minds frequently and have a tendency to forget facts.

External factors may introduce conflicting influences which affect the way in which a questionnaire is answered. Although men are well used to external influences and communicating with outsiders, women are cloistered and secluded from mixing with others outside the family. To maintain their privacy, females may alter their responses with repeated administrations of the questionnaire. Many female patients discussed personal and social problems with the nurse-interpreter and it is possible that they did not want to repeat their personal details to a second person and hence tried to diminish their feelings.

Physiological factors affect all members of the population, but the female students may have been more susceptible to these changes. The first test was completed when the students arrived at the site for a seminar and the second test was completed later during a break. Females have to travel together by bus while male students, who have their own cars, are more independent. Therefore, female students may suffer more from the effects of heat and dehydration, their symptoms improving with fluids and rest. This effect was also seen in those females attending the primary health care clinics who participated in the inter-rater reliability tests, where the responses may have improved after resting while waiting to see the second researcher. In any case, the difference in responses between males and females appears to be consistent. Finally, males offered more negative responses to fatigue. In general, negative responses tend to be more concordant and this may be another reason for the consistent male responses.

The Arabic fatigue scale was found to be valid, in that it measured what it was intended to measure — the severity of fatigue symptoms among patients in an Arabic population. The interpretation which is placed on these results, however, is also a part of the validation process. A questionnaire can be reliable and consistent, but be consistently measuring some effect other than that for which it was intended. Fatigue may overlap with other conditions, for example depression, so it possible that what is being measured is another construct with similar symptoms to fatigue. Therefore, the process of validation should be continuous, providing more evidence throughout the study that the questionnaire measures the intended construct. In this case, all those involved agreed on the validity of the translated questionnaire. In summary, the Arabic translation has been shown to be both reliable and valid, especially in the case of male respondents.

References

1. McWhinney IR. *A textbook of family medicine*, 2nd ed. New York, Oxford University Press, 1989.
2. Chalder T et al. Development of a fatigue scale. *Journal of psychosomatic research*, 1993, 37(2):147–53.
3. Lewis G et al. Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. *Psychological medicine*, 1992, 22(2):465–86.
4. Schoonjans F et al. MedCalc: a new computer program for medical statistics.

Computer methods programs biomedica, 1995, 48(3):257–62.

5. Hui C, Triandis HC. Measurement in cross-cultural psychology: a review and comparison of strategies. *Journal of cross-cultural psychology*, 1985, 16:131–52.

6. Eckes T. Features of men, features of women: assessing stereotypic beliefs about gender subtypes. *British journal of social psychology*, 1994, 33:107–23.

Corrections

1- Role of health education programmes within the Libyan community by A.A. Elfituri, M.S. Elmahaishi and T. H. MacDonald. EMHJ Vol. 5 No. 2 March 1999, page 268.

The list of authors in Arabic should read

عبد الباسط الفيتورى ومحمد المحيشى وثيودور ماكدونالد

2- Global and regional data on neuropsychiatric disorders. EMHJ Vol. 5 No. 2 March 1999, page 405.

Source should read *The World Health Report, 1999.* Geneva, World Health Organization, 1999: 106–7