UvA-DARE (Digital Academic Repository)

Validation of the EORTC QLQ-C30 quality of live questionnaire through combined qualitative and quantitative assessment of patient-observer agreement

Groenvold, M.; Klee, M.C.; Sprangers, M.A.G.; Aaronson, N.K.

DOI

[10.1016/S0895-4356\(96\)00428-3](https://doi.org/10.1016/S0895-4356(96)00428-3)

Publication date

1997

Published in

Journal of Clinical Epidemiology

[Link to publication](#)

Citation for published version (APA):

Groenvold, M., Klee, M. C., Sprangers, M. A. G., & Aaronson, N. K. (1997). Validation of the EORTC QLQ-C30 quality of live questionnaire through combined qualitative and quantitative assessment of patient-observer agreement. *Journal of Clinical Epidemiology*, 50(4), 441-450. [https://doi.org/10.1016/S0895-4356\(96\)00428-3](https://doi.org/10.1016/S0895-4356(96)00428-3)

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Validation of the EORTC QLQ-C30 Quality of Life Questionnaire Through Combined Qualitative and Quantitative Assessment of Patient–Observer Agreement

Mogens Groenvold,^{1,2,*} Marianne Carol Klee,²
Mirjam A. G. Sprangers,³ and Neil K. Aaronson⁴

¹DEPARTMENT OF SOCIAL MEDICINE, UNIVERSITY OF COPENHAGEN, DK-2200 COPENHAGEN, ²DEPARTMENT OF ONCOLOGY
ONK, NATIONAL UNIVERSITY HOSPITAL, DK-2100 COPENHAGEN, DENMARK, ³DEPARTMENT OF MEDICAL PSYCHOLOGY,
ACADEMIC HOSPITAL, NL-1105 AZ AMSTERDAM, AND ⁴DIVISION OF PSYCHOSOCIAL RESEARCH AND EPIDEMIOLOGY,
THE NETHERLANDS CANCER INSTITUTE, NL-1066 CX AMSTERDAM, THE NETHERLANDS

ABSTRACT. Patient-rated questionnaires are increasingly used to assess health-related quality of life. We studied one aspect of the validity of such measures that has rarely been investigated: do patients interpret questionnaires in the same way as do the researchers reporting the results? If not, there may be a problem. We employed the EORTC QLQ-C30 quality-of-life questionnaire to study 95 cancer patients and measured the agreement between (1) the patient's self-assessment and (2) an observer's rating of the patient's open-ended responses to the same questionnaire administered as an interview. The observer made qualitative recordings describing potential misinterpretations. The agreement between patients' and observers' ratings was high (median kappa = 0.85, range 0.49–1.00). The qualitative data revealed a few minor validity problems. One of these, selective reporting, may lead to systematic errors: some patients reported only what they considered "relevant" symptoms. The combination of quantitative and qualitative methods proved useful for questionnaire validation. J CLIN EPIDEMIOL 50;4:441–450, 1997. © 1997 Elsevier Science Inc.

KEY WORDS. Quality of life, validity, cancer, health status assessment, patient–observer agreement, questionnaire

INTRODUCTION

In recent years, the spectrum of endpoints used to evaluate medical treatments has widened. Physical, psychological, and social problems/symptoms related to the disease or its treatment are now to a greater extent recognized as important outcomes in cancer clinical trials. It is generally accepted that data concerning the patients' well-being should be provided by the patients themselves [1,2]. Standardized questionnaires for patient self-assessment have been developed and are used for that purpose in clinical research. It is critical that the validity and reliability of such measures be evaluated [2–4].

In this paper we address an issue that has rarely been investigated: whether the questions and response categories in self-assessment questionnaires have the same meaning for the patients and the researchers who interpret and report

the data. The validity of the questionnaire would be compromised if this were not the case.

In the self-assessment of health by means of a questionnaire, the patient reads and interprets the question, relates this perception to his or her situation, and transforms a more or less complex answer to one of the predefined response categories. Despite increasing involvement of patients in questionnaire development, health-care professionals often play a major role in formulating the questions. However, patients and health-care professionals do not necessarily share the same psychological and cultural frame of reference.

Belson [5] tested a set of survey questions such as "How many days of the week do you usually watch television?" and described numerous misinterpretations. On average, only 29% of the respondents interpreted a question exactly as intended. One of Belson's conclusions was that "... a great deal of communication failure arose from misinterpretation of words and phrases which are very commonly used in current question design" ([5], p. 351). Belson's method involved an extensive and detailed interview with the re-

*Address for correspondence: Mogens Groenvold, M.D., Ph.D., Department of Social Medicine, University of Copenhagen, Blegdamsvej 3, DK-2200 Copenhagen N, Denmark.

Accepted for publication on 5 November 1996.

spondents about their understanding of each question and was thus very time-consuming. It is unknown to what extent Belson's results from commercial survey research are relevant to the use of questionnaires in medical research. However, they illustrate the fact that misinterpretation of questions may be more common than expected.

Many researchers have compared patients' self-assessment on quality-of-life questionnaires with "proxy ratings" (made by health-care providers or significant others) [6]. Generally the concordance between patients and proxies is far from optimal [6]. One explanation of less than optimal patient-proxy agreement is that proxies have insufficient insight into the subjective state of the patient. Alternatively, it may be the case that questions and response categories in quality of life questionnaires are interpreted differently by patients and health-care professionals. In the latter case, it is important to determine whether the patients understand questions differently from what was intended and thus differently from the way in which they are later reported by the researchers.

In the current study, we employed the EORTC QLQ-C30 quality-of-life questionnaire, which was developed for patient self-assessment in cancer clinical trials [7-9]. Previous studies have found the instrument to be a valid and reliable measure of quality of life aspects relevant to lung cancer patients [8,9] and other cancer patient populations [10-14]. We have previously published item bias analysis of the EORTC QLQ-C30 when used with breast cancer patients. Evidence of item bias (differential item functioning) was found in specific applications [15]. The assumption that patients understand the questions and response categories of the EORTC QLQ-C30 as they were intended has not been studied systematically.

The primary objective of the current study was to determine whether patients interpret the questionnaire in the same way as the researchers who analyze and report the results and, if not, how these interpretations differ. Specifically, we have investigated the extent of agreement between (1) the patient's response to the EORTC QLQ-C30 and (2) an observer's rating of the patients' open-ended responses to the same questionnaire administered as an interview.

METHODS

Patients

BREAST CANCER PATIENTS. From February 1992 to April 1992, 57 patients were invited to take part in the study. All patients had been surgically treated for breast cancer 3-6 months earlier and were a random sample selected from approximately 2000 breast cancer patients taking part in a longitudinal quality-of-life study of adjuvant therapy (Danish Breast Cancer Cooperative Group program for adjuvant therapy and follow-up (DBCG-89) [16]). The patients selected were those scheduled to receive the second or third

questionnaire (three or five months after the initiation of adjuvant therapy, respectively).

GYNECOLOGICAL CANCER PATIENTS. From February 1992 to October 1992, 88 patients with cervical, endometrial, or ovarian cancer were invited to participate. Of these, 70 patients took part in a longitudinal quality-of-life study at the Department of Oncology of the National University Hospital in Copenhagen. All patients scheduled for a questionnaire (1-24 months after the initiation of primary radiotherapy or chemotherapy) were contacted. The remaining 18 patients were enrolled at follow-up after primary radiotherapy or chemotherapy at the Department of Oncology of Herlev University Hospital. This latter group of patients was included in order to approximate the size of the group of breast cancer patients.

Setting

A letter was mailed to the patients stating (1) the purpose of the study; (2) that a research nurse would contact the patient by telephone within a few days to ask whether the patient was willing to participate; and (3) that the enclosed questionnaire was not to be filled in before the nurse had called.

If the patient agreed to participate, the nurse arranged a day and time for a telephone interview. She asked the patient to fill in the questionnaire at home one hour before the interview and to place it in an envelope and seal it. If the one-hour interval was not feasible, an interval of up to 24 hr before the interview was acceptable.

Measures

QUESTIONNAIRES. The EORTC QLQ-C30 (version 1.0) [7-9] was used for all patients. This questionnaire has been developed by the European Organization for Research and Treatment of Cancer (EORTC) Study Group on Quality of Life to assess topics relevant to cancer patients. It is designed to be self-administered by the patients and consists of 30 items, 24 of which are organized into nine scales: Physical functioning, Role functioning, Emotional functioning, Cognitive functioning, Social functioning, Global health status/quality of life, Fatigue, Nausea and vomiting, and Pain. In the current study, however, the analysis was carried out at the individual item level.

The response categories of the EORTC QLQ-C30 are "No (1)"/"Yes (2)" (items 1-7), "Not at All (1)"/"A Little (2)"/"Quite a Bit (3)"/"Very Much (4)" (items 8-28), or 1-7 scales with 1 anchored to "Very poor" and 7 to "Excellent" (items 29 and 30). Item scores were linearly transformed to a range from 0 to 100 [9].

In the questionnaire for the breast cancer patients, the EORTC QLQ-C30 was followed by 39 additional items.

The questionnaire used for the gynecological cancer patients contained 73 additional items, some of which were inserted among the items of the EORTC QLQ-C30. Only the results concerning the EORTC QLQ-C30 are reported here.

INTERVIEW. The interviews were conducted by four nurses working in an oncological out-patient clinic. The nurses were trained in interviewing techniques and were given extensive feedback on their first interviews. The interviews were audiotaped. The nurses completed the questionnaire during the interview as an aid to the interviewing task. However, these interviewer-rated questionnaires were not analyzed. The following procedural guidelines were used: (1) Call the patient by telephone at the fixed time. (2) Ask when the patient filled in the questionnaire and record this. If the patient has forgotten to fill in the questionnaire arrange a new time for the interview. (3) Ask if the questionnaire has been put into the sealed envelope. (4) Explain the "rules" of the interview: the patient should try to answer without thinking of the response categories from the questionnaire and should not try to recall what she had answered when filling it in. (5) Read the questions aloud from the questionnaire exactly as they are written. (6) Ask the patient to elaborate her answer (by repeating the question, or by asking the meaning but not the wording of the response categories [e.g., "To what extent?"]) until you feel confident that you can tick the most appropriate response category. (7) If the patient uses a response category from the questionnaire in her answer, ask her to formulate her answer in a different way. (8) If the patient asks how a question should be understood then ask her how she understands it. (9) Interventions should be kept outside the context of the interview. If, during interviews, feelings or problems were raised that required more in-depth discussion, such discussions were deferred until the formal, structured part of the interview had been completed.

OBSERVER'S RATING. An observer filled in an identical questionnaire while listening to each audiotaped interview. As the aim of this study was to examine whether items had the same meaning for the patients and the researchers who interpret and report the data, the observers were the principal investigators of the clinical studies: one of the authors (MG) rated the breast cancer patients' interviews, while another author (MK) rated the gynecological cancer patients' interviews.

The questionnaires filled in by the patients were not available to the observers at this stage. Thus, the observers had no knowledge, while rating, of the extent to which their ratings were in agreement with those provided by the patients themselves. The guidelines were: (1) Check the response category that is most appropriate to the patient's answer *accepting the patient's norms and judgments but correcting misunderstandings*. For example, if the patient answered that

she only sleeps four hours per night and indicated that this is just a minor problem, this judgment should form the basis of the rating even if the rater would personally perceive it differently. On the other hand, if a clear misunderstanding was evident, for example that the patient had based her answer on the last month even if the question was restricted to the last week, the rater was to correct this by using the information regarding the last week. In other words, the observer should aim at "scoring correctly" in the sense of translating the substance of the patients' open-ended responses as faithfully as possible. If the patient revealed her initial (written) answer, the observer was to ignore that information. (2) Describe *potential* problems of interpretation in detail while rating the interviews. As the observer did not have access to the patient's self-rating, he or she could only guess as to when he or she was in agreement with the patient's original rating. The idea was to describe any *potential* (suspected) source of disagreement. This included cases where (a) the observer was in doubt about what to answer; (b) was in some way surprised by the patient's response; (c) believed he or she had identified a misunderstanding; or (d) the patient had realized during the interview that she had originally misinterpreted the question.

Statistical Analysis

Characteristics of the patients who did and did not agree to be interviewed were compared using Fisher's exact test or chi-square test (two-tailed).

Mean scores for each item of the EORTC QLQ-C30 were calculated. Patient and observer score distributions were compared using Wilcoxon signed ranks test (two-tailed) [17]. The level of significance was 0.05.

Agreement between the patients' ratings and the observer's ratings was estimated by calculating (1) overall agreement (all items), (2) the kappa coefficient (items 1–7), or (3) the weighted kappa coefficient (items 8–30).

(1) For each item, *overall agreement* was defined as the number of times the patient and the observer chose the same response category, divided by the number of times the item was answered by both the patient and the observer.

(2) *Kappa* [18] is a coefficient of agreement which is corrected for chance agreement. Values of kappa range from –1 to +1. A value of one corresponds to perfect agreement. Zero reflects that there is no more agreement than would be expected by chance, and a negative value means that the agreement is lower than one would expect by chance. Kappa was calculated for items 1–7 which have "No"/"Yes" response categories.

(3) Items 8–28 have ordinal response scales and items 29–30 have interval response scales. In the ordinary calculation of kappa all levels of disagreements are treated in the same way. With *weighted kappa* [19] it is possible to *graduate*

(weight) agreement/disagreement. Quadratic weights were used ([20], p. 118) yielding values identical to the intraclass correlation coefficient [21]. Standard error for kappas and weighted kappas was calculated according to the procedure provided by Fleiss *et al.* [22].

In the following sections of the paper kappa values and weighted kappa values will be termed "kappa(s)." The calculations of kappa and weighted kappa were programmed with the SAS computer package [23].

Landis and Koch [24] proposed categories for judging kappa values: kappa < 0.00 was termed "poor agreement," 0.00–0.20 "slight agreement," 0.21–0.40 "fair agreement," 0.41–0.60 "moderate agreement," 0.61–0.80 "substantial agreement," and 0.81–1.00 "almost perfect agreement." As a framework for analysis we decided *a priori* that kappa values equal to or less than 0.40 indicated potential problems, values up to 0.60 also deserved attention, whereas kappa values of 0.61 and above indicated acceptable results.

RESULTS

Patients

BREAST CANCER PATIENTS. Of the 57 patients invited to take part in the study, 46 (81%) were successfully interviewed. Eight patients declined participation. Interviews with 3 patients could not be arranged or were not effectuated for practical reasons. The mean age of the 46 patients interviewed was 51 years (standard deviation 9, range 29–70 years). Of these, 40 had undergone a mastectomy, and 6 had undergone a lumpectomy. Nine patients did not receive any adjuvant systemic treatment, 20 patients received chemotherapy (cyclophosphamide, methotrexate, and fluorouracil [CMF], or cyclophosphamide, epirubicin, and fluorouracil [CEF]), 8 were treated with tamoxifen, and 9 had undergone ovarian irradiation. Of the 46 patients, 37 were living with a partner, and 35 were employed full or part-time. No statistically significant differences in any of these background variables were found between those patients who were and who were not successfully interviewed.

GYNECOLOGICAL CANCER PATIENTS. Eighty-eight patients were invited to take part in the study, and 49 (56%) were successfully interviewed. Sixteen patients declined to be interviewed, while interviews with 23 patients could either not be arranged or were not effectuated for practical reasons (e.g., some of these patients had filled in the questionnaire before reading the accompanying letter stating that they should await a telephone call). The 49 interviewed patients had a mean age of 60 years (standard deviation 11, range 35–77 years). Of these, 28 had a cervical cancer diagnosis, 7 endometrial cancer, and 14 patients ovarian cancer. Forty-three patients were in the follow-up phase, and 6 were receiving radiotherapy or chemotherapy (cisplatin and ifosfamide). Twenty-nine women were living with a partner, and 21 were full or part-time employed.

Again, no statistically significant differences were found between those interviewed and those who were not successfully interviewed.

Quantitative Data

EORTC QLQ-C30 SCORES. Mean scores for ratings by patients and observers are presented in Table 1. In the questionnaires filled in by the patients, the proportion of missing values was 0–2% for 28 out of the 30 items. The two items with higher proportions of missing data were item 16 (constipation, 4%) and item 19 (pain, 12%). All missing values for item 19 were from the questionnaire completed by the gynecological cancer patients where this item was preceded by additional study-specific items on pain. Some patients reporting no pain skipped part of this section of items. No statistically significant differences were found between the patients' and observers' mean scores for any of the QLQ-C30 items.

OVERALL AGREEMENT BETWEEN PATIENTS AND OBSERVERS. In the total patient population, the median overall agreement was 0.85 (range 0.47–1.00). The value was above 0.80 for 21 of the 30 items (Table 2).

CHANCE-CORRECTED AGREEMENT BETWEEN PATIENTS AND OBSERVERS. The kappa values are shown in Table 2. In the total patient population, the median kappa was 0.85 (range 0.49–1.00). There was "almost perfect agreement" (kappa 0.81–1.00) for 18 items and "substantial agreement" (kappa 0.61–0.80) for 9 items. The remaining three items (nos. 3, 4, 28) were in the range termed "moderate agreement" (kappa 0.41–0.60). No items had kappas below this level. The mean difference between the paired kappas in the two sub-populations was 0.03 (standard deviation 0.18), the values for breast cancer patients being slightly higher than for gynecological cancer patients.

Qualitative Data

The observers rating the tape-recorded interviews made a total of 67 comments on 24 items. The majority (49) were made by the rater of the gynecological patients. The number of comments for any given item ranged from 0 to 7. The comments could be grouped into the following categories:

Category 1 (11 comments). The response was surprising to the observer, typically because it showed a very specific interpretation. It could, however, not be called a misinterpretation. Examples include the following: *Item 13:* The patient had had severe and continuous nausea but had been eating very frequently to try to reduce the nausea. She therefore answered definitely "No" to the question of whether she had "lacked appetite" even though her frequent eating was not related to hunger. *Item 19:* The patient was on sick leave because of pain. While at home, she could do her daily activities without pain. Her answer was that

TABLE 1. Mean scores* and standard deviations of the 30 items in the EORTC QLQ-C30. Breast cancer patients ($n = 46$) and gynecological cancer patients ($n = 49$)

Item content (abbreviated)	Breast cancer		Gynecological cancer	
	Patient Mean (SD)	Observer Mean (SD)	Patient Mean (SD)	Observer Mean (SD)
1. . . . any trouble doing strenuous activities, like carrying . . .	70 (47)	70 (47)	49 (51)	51 (51)
2. Do you have any trouble taking a <i>long</i> walk?	67 (47)	67 (48)	53 (50)	59 (50)
3. Do you have any trouble taking a <i>short</i> walk outside . . .	98 (15)	100 (0)	96 (20)	98 (14)
4. Do you have to stay in a bed or a chair for most of the day?	98 (15)	93 (25)	96 (20)	98 (14)
5. Do you need help with eating, dressing, washing yourself . . .	100 (0)	100 (0)	98 (14)	98 (14)
6. Are you limited in any way in doing job/household jobs?	72 (46)	67 (47)	75 (44)	71 (46)
7. Are you completely unable to work . . . or to do household job	100 (0)	100 (0)	96 (20)	96 (20)
8. Were you short of breath?	18 (27)	16 (27)	11 (23)	11 (22)
9. Have you had pain?	22 (27)	23 (31)	15 (24)	11 (18)
10. Did you need to rest?	40 (31)	40 (33)	29 (27)	26 (26)
11. Have you had trouble sleeping?	27 (28)	31 (37)	11 (18)	11 (20)
12. Have you felt weak?	23 (27)	22 (31)	22 (25)	15 (25)
13. Have you lacked appetite?	13 (25)	13 (26)	11 (21)	10 (20)
14. Have you felt nauseated?	22 (31)	21 (30)	9 (19)	10 (20)
15. Have you vomited?	6 (16)	8 (21)	4 (13)	4 (10)
16. Have you been constipated?	7 (15)	7 (20)	11 (20)	8 (16)
17. Have you had diarrhea?	9 (17)	9 (20)	16 (26)	14 (22)
18. Were you tired?	39 (32)	41 (35)	35 (32)	32 (32)
19. Did pain interfere with your daily activities?	11 (22)	10 (23)	9 (18)	5 (18)
20. . . . difficulty in concentrating on things, like . . .	82 (30)	80 (35)	86 (24)	88 (24)
21. Did you feel tense?	77 (33)	79 (33)	78 (28)	82 (24)
22. Did you worry?	67 (32)	69 (32)	76 (27)	82 (27)
23. Did you feel irritable?	79 (28)	82 (29)	84 (24)	86 (21)
24. Did you feel depressed?	72 (31)	71 (33)	76 (30)	78 (30)
25. Have you had difficulty remembering things?	80 (28)	77 (32)	79 (28)	80 (30)
26. . . . condition/treatment interfered with . . . <i>family</i> life?	88 (26)	89 (23)	93 (17)	94 (18)
27. . . . condition/treatment interfered with . . . <i>social</i> activities?	89 (23)	91 (24)	87 (23)	92 (20)
28. . . . condition/treatment caused you financial difficulties?	12 (25)	6 (20)	7 (18)	7 (21)
29. How would you rate your overall <i>physical</i> condition . . . ?	67 (25)	65 (28)	72 (25)	74 (17)
30. How would you rate your overall <i>quality</i> of life . . . ?	70 (28)	68 (29)	77 (23)	75 (18)

*The scores are linearly transformed. For items 1–7, 20–27, and 29–30 (functional items), a higher score represents a higher level of functioning. For items 8–19 and 28 (symptoms) a higher score represents a higher degree of symptoms.

pain had not interfered with her daily activities. *Item 24*: The patient had been unhappy but would not call herself “depressed”. She perceived “feeling depressed” as an “either/or” situation rather than as a question of degrees. Thus, while the patient reported that she was not depressed, the observer believed that the response “A little” might have been a more appropriate description.

Category 2 (12 comments). “Yes, I had the symptom, but it was due to something else” (i.e., not cancer or its treatment). Examples include: *Item 2*: The patient was unable to take a long walk due to urinary incontinence. She had been in doubt as to how to answer the question because this was “an old problem” and therefore not related to the present treatment. *Item 8*: “No, my dyspnea is due to weight gain, it has nothing to do with the disease.” *Item 9*: Five comments were related to pain that was attributed to other health problems (e.g., arthritis, coughing, etc.) that the patients believed were not relevant to the subject of the questionnaire. They therefore stated that this pain should not be taken into account. *Item 25*: “No, not more than I always

have. During the last couple of years it has become very difficult for me to remember things.”

Category 3 (28 comments). The patient was in doubt as to how to interpret the question or how to relate it to her situation. Examples include: *Item 1*: The patient could do strenuous activities but would be very tired afterwards. She had been in doubt as to which of the two response categories to choose. *Item 2*: The patient had not tried to take a long walk because—even though she would probably be able to do so—she knew that it would make her extremely tired afterwards. She had been in doubt whether to answer “Yes” or “No.” *Item 6*: Three patients could do all their usual household jobs, although it took longer. They were unsure if this meant that they had been “limited in any way in doing . . . household jobs?” *Item 12*: Five patients made spontaneous comments on how they perceived the relation between being tired and feeling weak. Two perceived weakness more as a trait (which they did not have) than as a state, and said that, even though they were very tired, they did not feel weak. The remaining three patients concluded

TABLE 2. Comparison of EORTC QLQ-C30 ratings by patients and observers. Overall agreement and kappa/weighted kappa values in breast cancer patients (BRC), gynecological cancer patients (GYN) and the two groups combined (TOTAL)*

Item	Overall agreement			Kappa/weighted kappa (SE)		
	BRC	GYN	TOTAL	BRC	GYN	TOTAL
1. Strenuous activities	1.00	0.94	0.97	1.00 (0.00)	0.87 (0.07)	0.93 (0.04)
2. Long walk	1.00	0.94	0.97	1.00 (0.00)	0.88 (0.07)	0.93 (0.04)
3. Short walk	0.98	0.98	0.98	0.00 (0.00)	0.66 (0.32)	0.49 (0.31)
4. Stay in a bed	0.96	0.98	0.97	0.48 (0.31)	0.66 (0.32)	0.56 (0.23)
5. Need help	1.00	1.00	1.00	^b	1.00 (0.00)	1.00 (0.00)
6. Limited doing job	0.96	0.94	0.95	0.90 (0.07)	0.84 (0.09)	0.87 (0.06)
7. Unable to work	1.00	1.00	1.00	^b	1.00 (0.00)	1.00 (0.00)
8. Short of breath	0.89	0.98	0.94	0.92 (0.04)	0.98 (0.02)	0.94 (0.02)
9. Pain	0.65	0.88	0.77	0.69 (0.08)	0.54 (0.19)	0.65 (0.09)
10. Rest	0.80	0.68	0.74	0.85 (0.06)	0.73 (0.08)	0.81 (0.04)
11. Trouble sleeping	0.70	0.88	0.79	0.84 (0.03)	0.81 (0.07)	0.85 (0.03)
12. Weak	0.78	0.66	0.72	0.81 (0.07)	0.66 (0.11)	0.75 (0.06)
13. Lacked appetite	0.91	0.96	0.94	0.92 (0.05)	0.95 (0.04)	0.93 (0.03)
14. Nauseated	0.93	0.98	0.96	0.96 (0.02)	0.97 (0.03)	0.97 (0.02)
15. Vomited	0.96	0.98	0.97	0.93 (0.03)	0.91 (0.06)	0.93 (0.03)
16. Constipated	0.96	0.96	0.96	0.80 (0.11)	0.81 (0.14)	0.81 (0.09)
17. Diarrhea	0.91	0.90	0.90	0.75 (0.11)	0.74 (0.16)	0.75 (0.11)
18. Tired	0.76	0.75	0.76	0.88 (0.04)	0.80 (0.07)	0.84 (0.04)
19. Interfere daily activities	0.87	0.84	0.86	0.78 (0.13)	0.64 (0.17)	0.73 (0.11)
20. Difficulty concentrating	0.83	0.86	0.84	0.91 (0.03)	0.81 (0.07)	0.87 (0.03)
21. Tense	0.87	0.67	0.77	0.93 (0.03)	0.73 (0.05)	0.85 (0.03)
22. Worry	0.74	0.78	0.76	0.86 (0.04)	0.62 (0.13)	0.76 (0.07)
23. Irritable	0.85	0.78	0.81	0.85 (0.07)	0.68 (0.09)	0.79 (0.06)
24. Depressed	0.89	0.82	0.85	0.90 (0.05)	0.88 (0.05)	0.90 (0.04)
25. Difficulty remembering	0.80	0.86	0.83	0.88 (0.04)	0.90 (0.04)	0.89 (0.03)
26. Family life	0.95	0.94	0.95	0.96 (0.03)	0.88 (0.07)	0.93 (0.03)
27. Social activities	0.89	0.82	0.85	0.89 (0.05)	0.64 (0.11)	0.77 (0.07)
28. Financial difficulties	0.80	0.90	0.85	0.36 (0.16)	0.66 (0.06)	0.49 (0.10)
29. Physical condition	0.50	0.46	0.48	0.75 (0.08)	0.67 (0.09)	0.72 (0.06)
30. Quality of life	0.46	0.48	0.47	0.79 (0.07)	0.69 (0.08)	0.76 (0.05)

*Items with kappas below 0.61 in the two groups combined are in bold face.

^bKappa cannot be calculated due to lack of variation.

that they felt weak because they were very tired. *Item 17*: The patient had to take several pills every day to treat her diarrhea, but could avoid the symptom as long as she kept up the treatment. She felt that it would give a wrong impression if she answered "No." *Item 19*: One patient who was in pain found the question difficult to answer because she did not think she had any "daily activities." Another patient perceived the question as difficult because she had no pain. *Item 30*: "Well, the problem is that I don't know what you mean by quality of life."

The problems reported with items 1, 2, and 6 are related to the fact that these items have dichotomous response categories, "No/Yes": these patients tended to make a finer distinction than allowed for in the questionnaire.

Category 4 (3 comments): Misunderstandings. Examples include: *Item 9*: "Oh no, definitely not." However, this patient suffered from substantial discomfort, but said that as long as she did not have to take analgesics it was not "pain." *Item 28*: Two patients realized during the interview that, in their initial rating of item 28 concerning financial difficulties, they had responded by stating the extent of expenses

due to the treatment or disease rather than financial difficulties. They had had no financial difficulties.

Category 5 (4 comments). The observer was in doubt as to which answer category should be checked. These comments described cases in which the observer found it difficult to respond due to insufficient information in the patient's reply (suggesting a deficiency in the interviewing). For example, for item 9, the patient described the location but not the extent of pain.

Category 6 (9 comments): Other comments. The majority of these comments described the patients' situation and motivation for a particular response. As an example: *Item 2*: The patient emphasized that she had not tried to take a long walk. She was, however, sure that it would cause her trouble.

DISCUSSION

This study was designed to evaluate the process of self-assessment associated with completion of a quality-of-life questionnaire commonly used for cancer clinical trials, the

EORTC QLQ-C30. The central question was whether cancer patients interpret the questions and response categories in the same way as do those analyzing and reporting the results.

We approached this question by assessing the level of agreement between the patients' self-assessments and observers' ratings of the same questionnaire items based on a detailed interview in which the questions were read aloud but the patient was requested not to use the standardized response categories. Qualitative recordings were made by the observer when potential problems of interpretation were encountered.

Discussion of Results

For most items of the EORTC QLQ-C30 the agreement between patients and observers was high: the median *overall agreement* was 0.85 (range 0.47–1.00), and the median *kappa* was also 0.85 (range 0.49–1.00). In the total patient population no items were below the *a priori* set limit of *kappa* (0.40), and 27 out of 30 items were above 0.60.

Compared with other studies (see [6] for an overview), the levels of agreement found in this study were relatively high. In contrast to many other studies we did not observe any cases of low agreement. This may have several explanations; the most important of which is probably differences in design. Our study was designed specifically to elucidate differences in perception. In other studies, differences may to a greater extent have reflected observers' insufficient knowledge about the patient's situation. Another explanation is the fact that the items of the EORTC QLQ-C30 have been through an extensive developmental process running in parallel in several languages, thereby reducing the ambiguity of the item wording. For example, the initial field-testing in the developmental phase required that each investigator "debrief" patients to determine whether there were questions that were upsetting, difficult to answer, confusing, etc. Such direct feedback from patients was used to refine the wording of items incorporated into the final version of the questionnaire.

Only 67 qualitative notations or comments were made by the observers. A total of 424 disagreements were noted. Comments may not always have been made in cases of disagreement. Thus, only a minority of the disagreements have been commented on. This illustrates that the qualitative and quantitative analyses are both needed to understand potential problems in a questionnaire.

The comments were grouped into six categories. Of these, category 2 ("Yes, I had the symptom, but it was due to something else") and category 4 (Misunderstandings) represent cases in which the observers were in disagreement with the patient about the interpretation of an item. The comments in the remaining categories increase our understanding of how patients interpret and respond to the items of the questionnaire but cannot be considered cases of disagreement between patients and observers as to the "correct" answers.

In general, the quantitative and qualitative results did not indicate any major validity problems with the questionnaire. In other words, the overall results suggest that patients and observers tend to interpret the questions and response categories in the same way.

In the following discussion, the items with kappas below 0.61 and those related to issues from the qualitative notes found to have a systematic character are examined more closely.

Items 1–7: Inspection of the examples in category 3 shows that the dichotomous ("No"/"Yes") response possibilities inherent in the questions themselves presented problems for some patients who wished to make finer distinctions than were allowed for in the questionnaire. The use of four response categories (as in items 8–28) would probably solve the problem. The dichotomous categories were originally chosen in order to generate a Guttman scale. The Guttman scaling was subsequently abandoned, and the index is now analyzed as a Likert scale [8]. A change from two to four response categories usually improves the reliability of a scale ([20], p. 35) and may improve its ability to detect clinically important change over time (i.e., responsiveness) [25]. The EORTC Study Group on Quality of Life is now testing a revised version of the questionnaire that includes four response options for these items.

Items 3 and 4 had relatively low *kappa* values despite very high overall agreement. This apparent paradox can be attributed to two aspects of the *kappa* statistic described by Feinstein and Cicchetti [26,27]. First, through the chance-correction mechanism, *kappa* is affected by prevalence. Only three patients answered "Yes" to items 3 or 4. Second, disparate agreement about positive and negative diagnoses (Yes/No) can be hidden in a large overall agreement because large disagreement about a "diagnosis" ("Yes") with low prevalence does not count very much in overall agreement. This was the case with items 3 and 4. The agreement about "No" was high (0.99 and 0.98, respectively) but the agreement about "Yes" was only 0.50 and 0.57, respectively (data not shown). Thus, there was agreement between patients and observers on the rating of *absence of these problems* but there were too few positive replies to allow evaluation of the agreement about the presence of these two problems.

Item 9 ("Did you have pain?"): The introductions to the quality-of-life study and to the questionnaire did not instruct the patients to consider the etiology of symptoms when responding. Nevertheless some patients reported symptoms only if they believed that they could be attributed to their cancer or its treatment. This was seen most frequently for item 9: the patients stated that they had pain, but that this should not be taken into account because it was "due to something else" (examples in category 2).

This phenomenon, which we label *selective reporting*, was also observed for five other items (category 2). Our impression is that selective reporting may be a general mechanism reflecting that patients are interested in the research issue and wish to contribute correct and useful data. It can be

seen as an example of a more general concept described in the behavioral science literature as “demand characteristics,” concerning the variety of cues stemming from a research setting that may affect a response [28].

Selective reporting may have at least two consequences. First, it may lead to systematic errors in the assessment of symptoms (e.g., underestimation of the true level or frequency of pain). Second, and possibly more problematic, in the comparison of groups of patients who vary in terms of their perception of the research in which they are participating, this effect may introduce *bias*. This is especially relevant when scores from patients taking part in a clinical trial are compared with scores from the general population (“norms”), where respondents may have no specific expectation as to what the investigators are looking for, and therefore may report their symptoms more completely. This could also be an issue in randomized clinical trials including a no-treatment arm. It may be possible to reduce selective reporting by modifying the instructions given to the patients. For example, by emphasizing that they should report all appropriate symptoms without considering their cause. However, the phenomenon must be explored in more detail before conclusions leading to actions can be drawn.

Item 12 (“Have you felt weak?”) was one of the items most frequently commented upon by the observers. The examples in category 3 show that some patients distinguished between being tired and feeling weak; the latter was suggested to be a trait rather than a state. The item may therefore be ambiguous. As it forms part of the “Fatigue Scale” (which also includes items 10 and 18), this ambiguity may lead to underestimation of the “true” level of fatigue.

Item 19 (“Did pain interfere with your daily activities?”): the observers’ comments (categories 1 and 3) suggest that this item (being a combination of two elements) may be more difficult to respond to than most other items. It appears likely that selective reporting could also occur for this pain item, but this was not recorded.

Item 28 (“Has your physical condition or medical condition caused you financial difficulties?”): in the example in category 4, two patients reported the extent of their expenses rather than any possible financial *difficulties*. An examination of the cases in which patients and observers disagreed on the rating of item 28 showed that in 8 of the 14 cases of disagreement the patient had answered positively (i.e., reporting financial difficulties) while the observer rated “Not at All.” These findings are compatible with the misreading suggested above having led some patients to report expenses rather than difficulties. The problem could probably be resolved by underlining “difficulties?”

GENERALIZING RESULTS. The validity of the findings of this study is supported by the similarity of the findings based on two patient groups interviewed by four different interviewers and rated by two different observers. This also suggests that the results may be generalized, at least to similar

patient populations, although we cannot rule out the possibility that other translations of the EORTC QLQ-C30 may contain wordings or meanings leading to other misinterpretations.

Discussion of the Research Design

In our opinion, the major strength of this study was that it combined a qualitative methodology allowing for prospective collection of data about possible sources of disagreement with a quantitative approach providing an estimate of the extent of agreement on each item.

It should be emphasized that we did not (and did not intend to) use the full capabilities of qualitative research methods. In-depth interviewing of the patients about the meaning of each item—by the methods described by Belson [5]—would provide more information but would also be much more time-consuming.

If the method presented here was used in earlier phases of questionnaire development or if lower levels of agreement were found, it would be relevant and probably highly informative to make selective audiotape transcripts of the cases of disagreement.

Even though the study (in contrast to earlier, purely quantitative studies) had a built-in mechanism for obtaining *possible explanations* of lack of agreement, we cannot know whether the reasons for disagreement identified by the observers were actually the major causes. The fact that the observers were unaware of the magnitude of agreement until all interviews had been rated, protected against retrospective *generation* of interpretations (only comments made during interviews were used in the analysis). However, this does not exclude the possibility that incomplete conclusions were reached during analysis of the data.

The magnitude of disagreement was intended to reflect possible divergence of interpretation between patients and observers. However, the following factors may have contributed to an over-estimation of disagreement: (1) Any measurement is associated with a certain degree of random error (lack of reliability). Thus, this type of attempt to identify systematic error by comparing two measurements is confounded by the (lack of) reliability of these measurements [14]. (2) The process of self-assessment when applied the first time may lead to a difference in perception. (3) The different situations in which responses were elicited (completion of a questionnaire by oneself versus in a social context with an unknown health-care professional) may affect the way in which the patient responds. (4) Despite the short time frame between assessments, the patient’s symptoms/problems (or the perception of these) may have changed.

Mechanisms tending to over-estimate agreement include: (1) An item may be misunderstood (or understood differently from the observer) by the patient without the observer noticing this. This may be the case if the observer does not pay sufficient attention or if the interviewer does not ask

the patient to elaborate her response in sufficient detail. (2) Patients may wish to reproduce the answers given in the first assessment (in order to appear consistent), and thus make efforts to provide an answer as close as possible to the first. However, this is only a source of error in cases where the patient had originally interpreted the item or response categories differently from the observer and this difference was not noted by the observer. (3) Patients may have revealed their initial (written) answer in the interview (despite instructions not to do this), and the observers may have used this information (despite the instruction not to use it).

The relative contributions of these potential sources of error cannot be estimated from this study. It is our impression that the factors tending to overestimate disagreement (especially the effect of reliability) are the more powerful, and that the method therefore provides a conservative estimate of the level of agreement.

Test-retest or inter-rater reliability studies can assess the reliability of the rating process. An inter-rater reliability study could have been built into the present investigation by having more than one observer rate each interview. However, when designing the study we believed that the data yielding the most clear-cut interpretation would come from comparing patient ratings with those of the (one) person responsible for the selection of items/questionnaires and the analysis and reporting of data. Adding extra observer ratings would obscure this principle. For the same reason, we did not include the questionnaires filled in by the interviewers (as an aid to the interviewing task) in the analysis.

Test-retest reliability of the EORTC QLQ-C30 was measured in a study of Norwegian cancer patients [14]. However, most of the results are not directly comparable with ours as Hjermstad *et al.* reported most data at the scale (not item) level and did not calculate kappas. Nevertheless, the overall agreement for the six items not forming scales can be compared. Interestingly, the overall (inter-rater) agreement for these six items obtained in the current study is higher than or identical to the (intra-rater) values found in the test-retest study.

At the scale level, Hjermstad *et al.* [14] reported Spearman correlation coefficients ranging from 0.70 to 0.90. Taken together, these results suggest that a considerable proportion of the disagreement we have found between patients and observer-raters can be attributed to the imperfect reliability of the questionnaire item response, rather than to diverging perceptions.

The method presented here may serve as a template that, perhaps with some minor modifications, may be adapted to a number of other situations in questionnaire development and validation. One possibility for modification of the design would be to change the order of self-assessments and interview, for example in a random half of the patients. We chose not to do this in order to have a self-assessment that was identical with that used in our clinical trials.

In conclusion, the results of this study indicated high levels of agreement between patients and observers about the rating of the EORTC QLQ-C30. This lends considerable support to the validity of cancer patient's self-assessment of their health-related quality of life by means of this questionnaire. A few minor problems of validity were identified. Knowledge of these, mainly the issue of selective reporting, may prevent future misinterpretations.

The combination of quantitative and qualitative methods proved very useful. In addition to its use in the validation of standardized questionnaires, the approach described here may be particularly useful in the development of new questionnaires.

The authors thank Nina Gyrtrup, Charlotte Bruun Knudsen, Irene Lindgreen, and Ann Raaberg for conducting the interviews, Lene Theil Skovgaard for statistical assistance, and Tage S. Kristensen, Henning T. Mouridsen, Jakob B. Bjørner, Jill Klee, and two anonymous reviewers for their comments on earlier drafts of this paper. The study was supported by grants from the Danish Cancer Society. At the time that this study was conducted, Mirjam Sprangers was on the staff of the Division of Psychosocial Research and Epidemiology, the Netherlands Cancer Institute, and was supported by a grant from the Dutch Cancer Society (NKI-90A).

Elements from this manuscript were originally proposed by M. Groenvold to K. Avlund and were used in K. Avlund et al.'s article, "Are Self-Ratings of Functional Ability Reliable?" (Scandinavian Journal of Occupational Therapy 1995; 2: 10-16). Unfortunately, the acknowledgement of M. Groenvold et al.'s contributions were not published in K. Avlund et al.'s article.

References

1. Slevin ML, Plant H, Lynch D, Drinkwater J, Gregory WM. Who should measure quality of life, the doctor or the patient? *Br J Cancer* 1988; 57: 109-112.
2. Bowling A. **Measuring Health: A Review of Quality of Life Measurement Scales.** Milton Keynes: Open University Press; 1991.
3. McDowell I, Newell C. **Measuring Health. A Guide to Rating Scales and Questionnaires.** New York: Oxford University Press; 1987.
4. Jaeschke R, Guyatt G. How to develop and validate a new quality of life instrument. In: Spilker B, Ed. **Quality of Life Assessment in Clinical Trials.** New York: Raven Press; 1990: pp. 47-57.
5. Belson WA. **The Design and Understanding of Survey Questions.** Hants: Gower; 1981.
6. Sprangers MAG, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: A review. *J Clin Epidemiol* 1992; 45: 743-760.
7. Aaronson NK, Bullinger M, Ahmedzai S. A modular approach to quality-of-life assessment in cancer clinical trials. **Recent Results Cancer Res** 1988; 111: 231-249.
8. Aaronson NK, Ahmedzai S, Bullinger M, Grabeels D, Estape J, Filiberti A, Flechtner H, Frick U, Hürny C, Kaasa S, Klee MC, Mastilica M, Osoba D, Pfausler B, Razavi D, Rofe PB, Schraub S, Sullivan M, Takeda F. The EORTC Core Quality-of-Life Questionnaire: Interim results of an international field study. In: Osoba D, Ed. **Effect of Cancer on Quality of Life.** Boca Raton: CRC Press Inc.; 1991: pp. 185-203.
9. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull

- A, Duez NJ, Filiberti A, Flechtner H, Fleishmann SB, de Haes JCJM, Kaasa S, Klee M, Osoba D, Razavi D, Rofo PB, Schraub S, Sneeuw K, Sullivan M, Takeda F. The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 1993; 85: 365–376.
10. Bjordal K, Kaasa S. Psychometric validation of the EORTC core quality of life questionnaire, 30-item version and a diagnosis-specific module for head and neck cancer patients. *Acta Oncol* 1992; 31: 311–321.
11. Niezgoda HE, Pater JL. A validation study of the domains of the core EORTC Quality of Life Questionnaire. *Qual Life Res* 1993; 2: 319–325.
12. Sigurdardottir V, Bolund C, Brandberg Y, Sullivan M. The impact of generalized malignant melanoma on quality of life evaluated by the EORTC questionnaire technique. *Qual Life Res* 1993; 2: 193–203.
13. Ringdal GI, Ringdal K. Testing the EORTC Quality of Life Questionnaire on cancer patients with heterogeneous diagnoses. *Qual Life Res* 1993; 2: 129–140.
14. Hjermstad MJ, Fossa SD, Bjordal K, Kaasa S. Test/retest study of the European Organization for Research and Treatment of Cancer Core Quality-of-Life Questionnaire. *J Clin Oncol* 1995; 13: 1249–1254.
15. Groenvold M, Bjorner JB, Klee MC, Kreiner S. Test for item bias in a quality of life questionnaire. *J Clin Epidemiol* 1995; 48: 805–816.
16. Danish Breast Cancer Cooperative Group. [DBCG-89. Program for Treatment and Follow-Up of Patients With Primary, Operable Breast Cancer] [in Danish]. Copenhagen: DBCG-Sekretariatet; 1989.
17. Siegel S, Castellan NJ, Jr. **Nonparametric Statistics for the Behavioral Sciences**. Singapore: McGraw-Hill; 1988.
18. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 20: 37–46.
19. Cohen J. Weighted kappa: Nominal scale disagreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968; 70: 213–220.
20. Streiner DL, Norman GR. **Health Measurement Scales—A Practical Guide to Their Development and Use**. Oxford: Oxford University Press; 1995.
21. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas* 1973; 33: 613–619.
22. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. *Psychol Bull* 1969; 72: 323–327.
23. SAS Institute Inc. **SAS Language: Reference, Version 6, First Edition**. Cary, NC: SAS Institute Inc.; 1990.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.
25. Guyatt GH, Townsend M, Berman LB, Keller JL. A comparison of Likert and visual analogue scales for measuring change in function. *J Chron Dis* 1987; 40: 1129–1133.
26. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990; 43: 543–549.
27. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 1990; 43: 551–558.
28. Orne MT. Demand characteristics and the concept of quasi-controls. In: Rosenthal R, Rosnow RL, Eds. **Artifact in Behavioral Research**. New York: Academic Press; 1969.