Consequences for caregivers of patients with severe mental illness: the development of the involvement evaluation questionnaire
van Wijngaarden, B.

Citation for published version (APA):
van Wijngaarden, B. (2003). Consequences for caregivers of patients with severe mental illness: the development of the involvement evaluation questionnaire

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: http://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.
Chapter 3

Caregiving in schizophrenia: Development, internal consistency and reliability of the European version of the Involvement Evaluation Questionnaire - EU Version (IEQ - EU)

Bob van Wijngaarden¹, Aart Schene², Maarten Koeter², José Luis Vázquez-Barquero³, Helle Charlotte Knudsen⁴, Antonio Lasalvia⁵, Paul McCrone⁶ and The EPSILON Study Group

¹Netherlands Institute of Mental Health and Addiction, Utrecht, The Netherlands
²Department of Psychiatry, Academic Medical Center, University of Amsterdam, Amsterdam
³Clinical and Social Psychiatry Research Unit, Department of Psychiatry, University of Cantabria, Santander, Spain
⁴Institute of Preventive Medicine, Copenhagen Hospital Corporation, Copenhagen University Hospital, Denmark
⁵Department of Medicine and Public Health, Section of Psychiatry, University of Verona, Italy
⁶Section of Community Psychiatry (PRiSM), Institute of Psychiatry, King's College London, UK

Published in British Journal of Psychiatry, 177 Isuppl. 39, s21-s27, 2000
Abstract

**Background:** In international research on the consequences of psychiatric illnesses for relatives of patients the need for an internationally standardised measure has been identified.

**Aims:** To test the internal consistency and the test-retest reliability of the Involvement Evaluation Questionnaire (IEQ) in five European countries.

**Method:** In the five collaborating sites the IEQ was administered twice to a sample of relatives or friends of patients with an ICD-10 diagnosis of schizophrenia. Reliability was tested using Cronbach's alpha, Intraclass Correlation Coefficients and Standard Error of Measurement. Reliability estimates were tested between sites.

**Results:** Test sample sizes ranged from 30 to 90 across sites and retest sample sizes ranged from 21 to 77. The distribution of scale scores differed across sites, which influenced the alpha values found. Cronbach's alphas of IEQ sub-scales and overall consequences score were substantial in most sites. On two occasions alpha values were moderate. Intraclass Correlation Coefficients were substantial to high in all sites. The Standard Errors of Measurement differed across sites, indicating differences in performance.

**Conclusion:** The reliability of the IEQ in five languages varies across sites, but is sufficiently high in at least four out of five sites. In one site reliability figures are lacking somewhat behind, so extra reliability research is advised there.

Introduction

Severe mental illness such as schizophrenia often imposes a considerable burden on the patients who suffer from it, as well as on their families and the wider society (Hatfield & Lefley, 1987). Patients' symptoms and their often poor personal and social functioning have a far-reaching impact on their own quality of life, while the nature of schizophrenia and its early onset often impoverish the lives and lifestyles of those who care for them. This issue has become even more important because of ongoing changes in the organisation of mental health care services: in particular, the shift from hospital-based to community-based services has resulted in some of the caring for (mostly) adults falling again on their family (or others involved).
When this is the case, normal and reciprocal caregiving between two (or more) adults changes into caregiving where one adult is dependent on the care of the other(s). The recipient of the care is disabled by a mental disorder with a long-term course, and for the caregiver(s), their caregiving role is out of synchrony with the appropriate stage of their own lifecycle (Schene, Tessler & Gamache, 1996). The consequences resulting from such a non-synchronised caregiving situation have for a long time been described as family or caregiver ‘burden’. However, this definition concentrates too much on the negative aspects of caring. Although mental disorders, particularly if they are long-term, disrupt family life, not all relatives experience their caring role as burdensome. Because of this we prefer the more neutral term ‘caregiver consequences’.

Research on the consequences of mental illness for patients’ relatives can be divided into three distinctive periods. First, starting in the 1950s, researchers described in detail all the different consequences for family members (Mandelbrot & Folkard, 1961; Wing et al., 1959). They paid particular attention to negative aspects. In the second period, starting in the early 1970s, family burden became one of the outcome measures in mental health service evaluation (Fenton, Tessler & Struening, 1979; Tessler, Killian & Tessler, 1980). Instruments were developed and used in studies that compared community approaches with the more classical approaches (Schene, Tessler & Gamache, 1994). In the third period, beginning in the early 1980s, interventions or treatment programs with a psycho-educational approach which aimed at a reduction of family burden, family stress or expressed emotion became the central point of interest (Kuipers & Bebbington, 1988). Recently, a fourth period has started, with the emphasis on relatives’ needs, perceptions and attributions (Barrowclough, Tarrier & Johnston, 1996; Scazuca & Kuipers, 1998), coping style (Budd, Oles & Hughes, 1998; Magliano et al., 1998) and mental health (Schene, Hoffmann & Goethals, 1998).

In 1994, Schene, Tessler & Gamache reviewed instruments measuring family or caregiver burden. They described 21 instruments mostly from English-speaking countries, of which 15 were developed in the 5-year period prior to the review, an indication of the growing importance of caregiver consequences. Since 1994, more new instruments have been developed, such as the Experience of Caregiving Inventory (Szmukler et al., 1996), the Perceived Family Burden Scale (Levine, Lancee & Seeman, 1996), and a generic instrument to assess the experience of caregiving (Schofield et al., 1997). Most of these instruments are interviews that can only be administered by interviewers, which make them time-consuming and expensive; in addition, they differ considerably in the number of items and domains covered. As far as we know, none of the above instruments has been translated from English into other languages, which limits their application, especially in Europe with its variety in languages. Therefore the European Psychiatric Services: Inputs Linked to Outcome Domains and Needs (EPSILON) Study also included the translation and validation of an
instrument to assess caregiver consequences (Becker et al., 1999). The instrument chosen was the Involvement Evaluation Questionnaire (IEQ: Schene & van Wijngaarden, 1992; Schene, Tessler & Gamache, 1996). The IEQ was chosen because [1] it was easy to administer, and [2] it is based on a variety of instruments developed in earlier years and covers a broad range of caregiver consequences (Schene, Tessler & Gamache, 1994).

Involvement Evaluation Questionnaire

Development of the IEQ: a brief history

The development of the IEQ started in 1987. In a randomised controlled trial, comparing in-patient and day patient treatment, an instrument was needed which could measure caregiver consequences (Schene et al., 1993). Since no such instrument existed in The Netherlands we developed one ourselves, starting with an extensive review of all empirical studies on family burden (Schene, 1986; Schene, Tessler & Gamache, 1994). We were then able to define the following desirable characteristics for such a research instrument: [1] it should be a questionnaire, covering all important domains, [2] it should be valid, reliable, easy to understand, and not time-consuming, and [3] it should cover a limited time frame and be sensitive to changes.

Items collected in the review of literature and existing instruments formed the basis of an IEQ item pool (Schene, Tessler & Gamache, 1996). This item pool was further extended with items emerging from interviews with professionals. A series of draft versions were piloted, and adapted if necessary. Since the principal aim in developing the IEQ was a reliable measure which would be sensitive to change, items relating to stigma, guilt, social network loss, suicide attempts by patients and other events that either happen rarely or are not sensitive to change were dropped.

The first version of the IEQ was used in four Dutch studies conducted between 1987 and 1990: [1] a comparative study of day treatment versus in-patient treatment (N=80), [2] a study among patients who had recently attempted suicide (N=80), [3] a study at the psychiatric department of a general hospital (N=80), and [4] a study among acute psychiatric patients in a community mental health centre (N=30). A psychometric analysis of these data and an updated literature review (in particular with regard to depression and families) resulted in the construction of a second Dutch version in 1992, which in the same year was translated into English, followed by translations in Portuguese, Finnish and German. These translations, however, did not follow the
procedures used in the EPSILON study (see Knudsen et al., 1999) and therefore have a different status from the translations used in the EPSILON study.

Structure and item content of the IEQ

The IEQ is a 31-item questionnaire, which is completed by the caregiver. The items relate to the encouragement and care that the caregiver has to give to the patient, to personal problems between patient and caregiver, and to the caregiver’s worries, coping and subjective burden. All items are scored on 5-point Likert scales (never, sometimes, regularly, often, always). Carers who have less than one hour’s contact with the patient during the 4 weeks previous to the completion of the instrument, skip items that refer to actual help and encouragement, because those items are considered to be not applicable. A total of 27 items can be summarised in four distinct sub-scales: (1) tension (nine items), which refers to the strained interpersonal atmosphere between patient and relatives, (2) supervision (six items), which refers to the caregiver’s tasks of guarding the patient’s medicine intake, sleep and dangerous behaviour, (3) worrying (six items), which covers painful interpersonal cognitions, such as concern about the patient’s safety and future, general health and health care, and (4) urging (eight items), which refers to activation and motivation; for instance, stimulating the patient to take care of her/himself, to eat enough and to undertake activities. In addition, a 27-item overall consequences score can be computed. The items and sub-scales are presented in Appendix III. In this paper the reliability testing of the sub-scales and overall score will be described.

For research purposes, the IEQ is normally extended with six additional modules. In the EPSILON study the following modules were used: (1) 15 socio-demographic and contact variables, such as age, gender, household composition, and amount of contact between patient and respondent, (2) eight items on extra financial expenses incurred on behalf of the patient, (3) three items on the caregiver’s use of professional help, (4) 11 items on the consequences for the patient’s children, (5) one open question for comments and additions, and (6) the 12-item General Health Questionnaire as a measure of caregiver distress (Goldberg & Williams, 1988; Dutch translation: Koeter & Ormel, 1991; Spanish translation: Gaite, 1997; Danish translation: Nielsen, undated; Italian translation: Servizio di Psicologia Medica Verona, undated).

The entire set takes about 20-30 minutes to complete; the IEQ module itself about 10 minutes. The IEQ can also be administered as a structured (telephone) interview. The IEQ can be used as both a research and a clinical instrument. Different ways of scoring are recommended, depending on the use of the IEQ: for research purposes the average scale scores based on the 5-point Likert scales are used for the computation of correlations with other instruments, in
clinical use, where average scale scores are not easy to interpret, item scores are dichotomised to 'no consequences' (categories 'never' or 'sometimes' on the Likert scale), and 'real consequences' (categories 'regularly', 'often', or 'always' on the Likert scale). In this case the scale scores directly reflect the number of consequences that are experienced. Also, major changes in consequences can easily be detected when an item score changes from 'no consequences' to 'real consequences', or the reverse. Collected data can be interpreted on both sub-scale as item level.

Validity, reliability, and applicability of the Dutch version of the IEQ

The 1992 version of the IEQ was tested in two Dutch studies, one among 680 members of an organisation of relatives of patients with psychotic disorders (Schene & van Wijngaarden, 1995), and one among 260 relatives of patients with affective disorders (van Wijngaarden, Schene & Koeter, 1996). From the results of both studies it was concluded that the IEQ adequately covers all major domains of caregiver consequences. The four identified sub-scales were obtained by factor analysis and cover the caregiver consequences of both psychotic and affective disorders.

The reliability of the IEQ proved to be satisfactory in the Dutch samples. The internal consistency (Cronbach's alpha) ranges from .74 to .85 for the four sub-scales to .90 for the overall score, test-retest effects were not found, and there were indications that the IEQ is sensitive for change (van Wijngaarden, Schene & Koeter, 1996). Validity was also satisfactory. The construction process as described above secured the content validity of the IEQ. This validity was confirmed by a qualitative analysis of the open question # 81, in which respondents were asked to add any issue that bothered, stressed or satisfied them in their relationship to the patient that was not already covered by the IEQ. Analysis of the replies to this question in about 1000 questionnaires did not reveal missing domains or variables (Schene & van Wijngaarden, 1993). In addition, separate analyses on the data regarding relatives of psychotic or depressed patients revealed factor structures that were very comparable with that of the combined sample. This consistency can also be considered as an indicator of content validity.

In an analysis of the relation between (1) caregiver consequences and characteristics of the patient, the caregiver and their relationship, and (2) caregiver consequences and caregiver distress, it was found that caregiver consequences were related to the patient's symptomatology and amount of time spent together. A path analysis revealed that caregiver consequences measured with the IEQ explained a substantial part of the relation between caregiver's distress and the patient, caregiver and relationship characteristics.
(Schene, van Wijngaarden & Koeter, 1998), emphasising the relevance of the concept.

Finally, the applicability of the IEQ proved to be good. Response rates were high, ranging from 70% to 81% (mailed survey with one reminder). The quality of the response was also high. Of 960 completed questionnaires, only 25 (2.6%) could not be used due to missing values.

Reliability of the IEQ

Sample

The IEQ was completed by relatives (or other significant persons) of patients with an ICD-10 diagnosis in Amsterdam, Copenhagen, London, Santander and Verona. For details on catchment areas and inclusion criteria see particular papers (Becker et al., 1999, 2000); Kastrup, 1998; Schene, Hoffmann & Goethals, 1998; Tansella et al., 1998; Thonicroft & Goldberg, 1998; Vázquez-Barquero & Garcia, 1999).

All patients who entered the study were asked to name a relative or other significant person who could be asked to complete the IEQ. However, not all of them were able to indicate someone who could complete the IEQ, either because they were not in close contact with others or because they refused to cooperate. The number of patients in each site included in the study varied from 52 to 107, with a total of 404. In Amsterdam and Copenhagen, where more patients live alone than in the others sites, the attrition was highest (about 40%). The number of respondents who completed the IEQ ranged from 30 (Copenhagen) to 78 (Santander), with a total of 285, which means an average response of about 70%. The number of retests ranged from 21 (Copenhagen) to 73 (London).

In the reliability protocol it was stated that at least 50 test-retest sets should be necessary for reliability testing. Once it became clear that the Amsterdam numbers would be lower, it was decided to do an extra sampling among 100 members of the Dutch organisation for relatives of schizophrenia patients. This sampling resulted in 52 test and 47 retest assessments, bringing the Amsterdam figures to a total of 90 tests and 77 retests.
Translation and cultural validation

The translation of the IEQ into the other languages largely followed the protocol described in Knudsen et al., 2000. This protocol included: (1) a translation into the four target languages by professional translators who were informed on the content of the IEQ; (2) a discussion of this translation by the translator and the research group, leading to a revision and a list of disputed items; (3) a back-translation in Dutch by a native speaker, who also gave his/her comments on the first translation and the disputed items; (4) a comparison of the back-translation with the original IEQ, discussed by the first translator and the researchers, leading to a second revision and list of disputed items; (5) a discussion of this revision in focus groups; (6) a discussion of the focus group result by the researchers and one of the translators, and (7) a third revision leading to the final version.

The focus group method is an arranged communication session among a selected group of persons who represent different parties involved. In case of the IEQ, they were representatives of patients, relatives, professionals and researchers. The conditions under which the focus group took place were well defined, including the role of the moderator, the physical setting, psychological climate, selection of subjects, instrumentation and aspects of data collection and analysis (Knudsen et al., 2000).

In the IEQ focus groups the translation and content of the instrument was discussed, with special emphasis on linguistic problems, the applicability and relevance of items, redundancy, and missing items. It was concluded that the instrument covers the domain of family burden. There were some problems with the response categories and items regarding education, type of professional help, income categories, and use of illegal drugs. The IEQ was adjusted in accordance with these comments (Knudsen et al., 2000).

Methodology reliability study

For the reliability testing of the IEQ two measures were used: (1) Cronbach's alpha for the internal consistency of the IEQ sub-scales, and (2) the Intraclass Correlation Coefficient (ICC) to estimate the test-retest reliability of the sub-scales. These estimates were computed for each site separately, and subsequently tested for inter-site differences. Since the reliability estimates are dependent on true score variance, inter-site comparisons will be affected if the variances are too different. In case of differences in score distribution, thereafter, the Standard Errors of Measurement ($SE_M$) were computed. These $SE_M$ scores, which are independent from the true score variance, were computed in two ways, either using Cronbach's alpha or the ICC in the formula (Schenk et al., 2000). In addition, pooled reliability estimates were computed to assess overall
reliability. As was pointed out by Schene and colleagues (2000) these pooled estimates are influenced by differences in score distribution and differences in reliability between sites. Pooled estimates should be treated with caution, especially where reliabilities are generally not very high, or lower in one or more sites.

The following analysis scheme was used:
1. Test on inter-site differences in score distribution (mean and variance; Anova and Levene test);
2. Assessment of the site-specific reliability estimates (Cronbach’s alpha, ICC, \(SE_M\)); the benchmark for substantial reliability was set to 0.70;
3. Test on inter-site differences in reliability estimates;
4. Reliability estimates from pooled data.

The reliability methodology of the EPSILON study and the computer programs used are described in detail by Schene and colleagues (2000).

Results

Score distributions

Table 3.1 presents means and standard deviations of the five samples, together with the test on homogeneity of these variances. It shows that both means and variances differ between sites. Mean scale scores and variances are generally high in Verona and low in Copenhagen. In all cases means differ significantly, and in three cases (supervision, worrying and urging) this also holds for the variances. As the 95% confidence intervals show, the contrasts are mainly between Copenhagen on one side and Santander and Verona on the other.

Internal consistency

Cronbach’s alphas for the IEQ sub-scales are presented in Table 3.2. The alpha values range from 0.68 to 0.86 for the sub-scales and from 0.87 to 0.91 for the overall score. In two cases the benchmark for substantial reliability was just not reached, the alpha for ‘supervision’ in London and that for ‘urging’ in Santander both having a value of 0.68.

Alpha testing between sites showed that on three sub-scales and the overall score the alpha values do not differ. Only on the sub-scale ‘urging’ were differences significant. The alpha in Santander is lower than those in
Amsterdam, London and Verona. Copenhagen also showed a lower alpha than London.

Differences in alpha values may be due to differences in score distribution. For that reason Table 3.2 also gives the $SE_M$ values. The $SE_M$ is lowest for all scales in Copenhagen. On the other hand, Santander and Verona showed relatively high $SE_M$ scores. This means that the IEQ seems more precise in Copenhagen and less precise in Santander and Verona.

As the $SE_M$ scores suggest, the differences in alpha values between sites are caused as much by differences in sample variance as by differences in true reliability. For instance, in Copenhagen alpha values are somewhat lower because of the low sample variance (Table 3.1), and in Verona it is just the other way around. Where a lower alpha is combined with a higher $SE_M$, reliability seems a bit problematic. This is the case for the sub-scale 'urging' in the Santander sample. Here the lowest alpha value is combined with the highest $SE_M$. Such a combination is not found in any other case. In the case of pooled data, all reliability estimates are substantial, ranging between 0.77 and 0.90.

**Test-retest reliability**

Table 3.3 presents the results of the test-retest reliability analysis. In all but one case, reliability is substantial to high (at least 0.70). The ICC values prove to be highest in Amsterdam, Copenhagen and London, demonstrating good reliability. In Verona, one ICC value (worrying) is only moderate. Although all reliabilities are substantial, in Santander the ICCs for 'supervision', 'worrying' and 'urging' are somewhat lower compared with the very high values in the other sites (see shaded area in Table 3.3). The $SE_M$ values indicate that the somewhat lower reliability of the IEQ in Verona and Santander can be attributed, at least partly, to higher measurement error. All differences between sites proved to be significant. The ICCs of the pooled data sets were all fairly high, ranging between 0.83 and 0.90.

The comparison of test and retest scores showed that retest scores in general were somewhat lower. In seven out of 25 cases these differences were significant, four times involving the sub-scale 'worrying', twice the overall score, and once 'urging'.

The length of the interval between test and retest may influence reliability. The IEQ does not measure a stable characteristic, but consequences for relatives, which may change over time. Thus, the longer the interval between test and retest, the higher the probability that the situation has changed.
Table 3.1 IEQ-EU sub-scales in the pooled sample and by site

<table>
<thead>
<tr>
<th>Sub-scale</th>
<th>Pooled n = 278</th>
<th>Amsterdam n = 65</th>
<th>Copenhagen n = 25</th>
<th>London n = 54</th>
<th>Verona n = 57</th>
<th>Santander n = 77</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>sd</td>
<td>mean</td>
<td>sd</td>
<td>mean</td>
<td>sd</td>
</tr>
<tr>
<td>Tension</td>
<td>8.3</td>
<td>3.8</td>
<td>7.7</td>
<td>3.2</td>
<td>7.1</td>
<td>2.0</td>
</tr>
<tr>
<td>Supervision</td>
<td>15.6</td>
<td>6.3</td>
<td>14.3</td>
<td>4.3</td>
<td>12.3</td>
<td>3.4</td>
</tr>
<tr>
<td>Worrying</td>
<td>15.4</td>
<td>6.4</td>
<td>14.1</td>
<td>5.6</td>
<td>12.7</td>
<td>3.6</td>
</tr>
<tr>
<td>Urging</td>
<td>50.6</td>
<td>16.3</td>
<td>46.7</td>
<td>14.7</td>
<td>41.3</td>
<td>9.4</td>
</tr>
<tr>
<td>Overall score</td>
<td>18.3</td>
<td>7.2</td>
<td>17.0</td>
<td>6.1</td>
<td>15.7</td>
<td>4.9</td>
</tr>
</tbody>
</table>

*Samples quoted for Overall score, sample sizes varied between 278 (Overall score) and 335 (Worrying)."
Table 3.2 Internal consistency of the IEQ-EU: alpha coefficients (95% CI) and SE in the pooled sample and by site

<table>
<thead>
<tr>
<th>Sub-scale</th>
<th>Amsterdam</th>
<th>Copenhagen</th>
<th>Santander</th>
<th>London</th>
<th>Verona</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tension</td>
<td>287/81(77-88)</td>
<td>68/54(38-73)</td>
<td>55/69(48-76)</td>
<td>77/80(71-83)</td>
<td>62/77(61-82)</td>
</tr>
<tr>
<td>Supervision</td>
<td>285/77(73-81)</td>
<td>69/70(60-71)</td>
<td>54/75(56-78)</td>
<td>77/74(69-81)</td>
<td>70/75(64-82)</td>
</tr>
<tr>
<td>Worrying</td>
<td>335/84(77-85)</td>
<td>88/81(70-90)</td>
<td>70/83(66-83)</td>
<td>78/83(70-88)</td>
<td>83/89(76-89)</td>
</tr>
<tr>
<td>Urging</td>
<td>291/79(73-80)</td>
<td>70/75(62-79)</td>
<td>25/71(45-86)</td>
<td>55/71(80-91)</td>
<td>28/47(28-82)</td>
</tr>
<tr>
<td>Overall score</td>
<td>278/90(88-93)</td>
<td>90/87(76-93)</td>
<td>65/77(54-89)</td>
<td>77/87(72-91)</td>
<td>57/64(50-73)</td>
</tr>
</tbody>
</table>

* Differences were found between Amsterdam – Santander *p* < 0.001, London – Santander *p* < 0.001.

* Differences were found between Amsterdam – Copenhagen *p* < 0.05, Santander – Verona *p* < 0.05.

52
<table>
<thead>
<tr>
<th>Sub-scale</th>
<th>Pooled n = 198</th>
<th>Amsterdam n = 65</th>
<th>Copenhagen n = 25</th>
<th>London n = 54</th>
<th>Santander n = 77</th>
<th>Verona n = 57</th>
<th>Test of equality of ICCs (P value)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICC</td>
<td>SE&lt;sub&gt;M&lt;/sub&gt;</td>
<td>ICC</td>
<td>SE&lt;sub&gt;M&lt;/sub&gt;</td>
<td>ICC</td>
<td>SE&lt;sub&gt;M&lt;/sub&gt;</td>
<td>ICC</td>
</tr>
<tr>
<td>Tension</td>
<td>.89</td>
<td>1.71</td>
<td>.92</td>
<td>1.21</td>
<td>.95</td>
<td>0.71</td>
<td>.97</td>
</tr>
<tr>
<td>Supervision</td>
<td>.83</td>
<td>1.54</td>
<td>.87</td>
<td>1.05</td>
<td>.98</td>
<td>0.50</td>
<td>.97</td>
</tr>
<tr>
<td>Worrying</td>
<td>.84</td>
<td>2.43</td>
<td>.87</td>
<td>0.87</td>
<td>.93</td>
<td>1.07</td>
<td>.98</td>
</tr>
<tr>
<td>Urging</td>
<td>.89</td>
<td>2.03</td>
<td>.93</td>
<td>1.39</td>
<td>.80</td>
<td>1.55</td>
<td>.98</td>
</tr>
<tr>
<td>Overall score</td>
<td>.90</td>
<td>5.07</td>
<td>.94</td>
<td>3.47</td>
<td>.93</td>
<td>2.37</td>
<td>.99</td>
</tr>
</tbody>
</table>

ICC: Intraclass correlation; SE<sub>M</sub>: standard error of measurement (square root of error components of variance)

<sup>1</sup> Samples quoted for Overall score; samples sizes varied between 198 (Overall score) and 261 (Worrying)
However, as was stated in the general reliability paper (Schene et al., 2000) a short time interval also may produce biased reliability estimates, due to the effect of memory. In the EPSILON study the interval was set at 1-2 weeks, but the actual average interval varied considerably between sites. In Santander this interval was 6.3 days (s.d.=2.6), in London 10.0 days (s.d.=8.7), in Amsterdam 11.6 days (s.d.=8.2), in Copenhagen 14.9 days (s.d.=5.4) and in Verona 22.4 days (s.d.=14.6). The somewhat lower reliabilities were found in the sites where time intervals were longest or shortest, so that there appears to be no direct connection between interval and reliability.

Discussion

The internal consistencies of the IEQ scales in general turned out to be satisfactory. In Amsterdam, Copenhagen and Verona all alpha values were substantial, while in both London and Verona, alpha was moderate in only one case. The comparison of alpha between sites was hindered by the differences in score distribution. As stated earlier, comparison of Cronbach’s alphas, and Intraclass Correlation Coefficients directly reflects measurement error, if it is assumed that the true score variances are comparable. The differences in score distribution, with Santander and Verona having the highest means and largest variances, as opposed to Copenhagen, have certainly influenced the test results. These differences seem to have caused the Santander and Verona alphas to be higher than the Copenhagen alphas. The Standard Errors of Measurement show that when the effect of variance differences is cancelled out, at face value the Copenhagen data seem most precise, and the Santander and Verona data less precise. A relatively low alpha combined with a high $SE_M$ is an indication of a somewhat lower performance of the instrument itself, rather than a characteristic of the sample on which the IEQ was tested. Considering this, the Santander alpha score on the sub-scale ‘urging’ seems to be a bit problematic. It is recommended that the reasons why only a moderate reliability estimate was found here should be explored further.

The estimates from pooled data were all substantial to high. Only in the case of the sub-scale ‘urging’ should one be careful in the interpretation of these findings, due to the significant differences between sites. In all other cases, no differences were found and the pooled reliability estimates can be considered valid.

The test-retest reliability estimates were all substantial, but differed between sites. ICC values were highest in Amsterdam, Copenhagen and London. Although ICC values are affected by differences in sample variance, the lower reliabilities tended to be those with high $SE_M$ values, and a low ICC combined
with high $SE_M$ may be an indication of lower performance. In Verona and Santander, the $SE_M$ values did indeed appear to be higher than at the other sites (although the differences have not been formally tested).

Whether the differences in reliabilities are caused by cultural differences, sampling, or test effects is not yet known. It was found that there was not a direct connection between the test-retest interval and reliability. The way in which the IEQ was administered might yield a possible explanation. In Verona and Santander, the IEQ was administered regularly as an interview or completed under the supervision of a research assistant, while in the three other sites, most IEQs were completed by the respondents themselves. As the IEQ is designed to be a self-administered questionnaire, an interview might add some extra bias. Additional data on self-administered IEQs will be necessary to investigate this hypothesis. Finally, in Santander and Verona more patients live with other people (parents, relatives). Because of the more intense contact between patient and relative in these situations, real changes – even in a short time – will be detected earlier than in the case in which the relative does not live with the patient.

Retest values were generally slightly lower than test values, indicating a certain test-retest effect. The absolute differences, however, were rather small, and were no higher in one site compared to another. This systematic test-retest effect does not explain the lower ICC scores in Verona and Santander.

Pooled data analyses resulted in rather high reliability estimates. As was stated earlier in this paper, these pooled estimates should be treated with caution, because ICC values are somewhat lower in two sites. On the other hand, the pooled estimates are sufficiently high and, combined with the fact that the lowest values are either moderate or in one case just below moderate, it is reasonable to conclude that overall reliability is good.

In summary, although the differences in sample variance make an exact test of the scales somewhat difficult, in general the IEQ scales have a substantial reliability in all sites. This conclusion is supported by the results of a Simultaneous Component Analysis in which factor analyses on all five separate data sets were compared in one single analysis. As it turned out in all sites very similar factors were found, indicating that the IEQ scales sufficiently cover all five samples.

The IEQ test-retest reliability analyses have been conducted on rather small samples, especially in Copenhagen (N=21). Samples in the other sites ranged from 47 to 77. In sites with larger samples, reliability issues could be studied in more detail. Bearing this in mind, one can conclude that despite some questions that still have to be answered, the reliability of the IEQ in the five
Development of the European version of the IEQ

EPISODE sites seems to be good enough for the moment to encourage the use of the instrument in European research. In doing so, larger data sets can be produced to study the validity and reliability of the IEQ in greater detail.

Acknowledgements

The following colleagues contributed to the EPISODE Study: Amsterdam: Maarten Koeter, Karin Meijer, Marcel Monden, Aart Schene, Madelon Sijsenaar, Bob van Wijngaarden. Copenhagen: Helle Charlotte Knudsen, Anni Larsen, Klaus Martiny, Carsten Schou, Birgitte Welcher. London: Thomas Becker, Jennifer Beecham, Liz Brooks, Daniel Chisholm, Gwyn Griffiths, Julie Grove, Martin Knapp, Morven Leese, Paul McCrone, Sarah Podfield, Graham Thornicroft, Ian R White. Santander: Andres Arriaga Arrizabalaga, Sara Herrera Castanedo, Luis Gaite, Andrés Herran, Modesto Perez Retuerto, José Luis Vázquez-Barquero, Elena Vázquez Bourgon. Verona: Francesco Amaddeo, Giulia Bisoffi, Dorian Cristofalo, Rosa Dall'Agnola, Antonio Lasalvia, Mirella Ruggeri, Michele Tansella. We would also like to acknowledge the sustained and valuable assistance of the users, carers and the clinical staff of the services in the five study sites.

This study was supported by the European Commission BIOMED 2 Programme (Contract BMH4-CT95-1151). In Amsterdam, the EPISODE study was partly supported by a grant from the Nationaal Fonds Geestelijke Volksgezondheid and a grant from the Netherlands Organisation for Scientific Research (940-32-007). In Santander the Epsilon Study was partially supported by the Spanish Institute of Health -FIS- (FIS Exp. N° 97/1240). In Verona additional funding for studying patterns of care and costs of a cohort of patients with schizophrenia were provided by the Regione del Veneto, Giunta Regionale, Ricerca Sanitaria Finalizzata, Venezia, Italia (Grant No 723/01/96 to Professor M. Tansella).
References


Gaite, L. (1999). *Cuestionario de salud general (IGHG-12 Spanish translation).* Santander: University of Cantabria, Clinical and Social Psychiatry Research Unit, Department of Psychiatry


Niezen, H. (undated). *General Health Questionnaire (IGHQ-12 Danish translation).* Odense: University Hospital, Department of Neuroradiology


