Methodology of a multi-site reliability study

EPSILON Study 3

AART H. SCHENE, MAARTEN KOETER, BOB VAN WIJNGAARDEN, HELLE CHARLOTTE KNUDSEN, MORVEN LEESE, MIRELLA RUGGERI, IAN R. WHITE, JOSÉ LUIS VÁZQUEZ-BARQUERO and the EPSILON STUDY GROUP

Background The European Psychiatric Services: Inputs Linked to Outcome Domains and Needs (EPSILON) Study aims to produce standardised versions in five European languages of instruments measuring needs for care, family or caregiving burden, satisfaction with services, quality of life, and socio-demographic and service receipt.

Aims To describe background, rationale and design of the reliability study, focusing on reliable instruments, reliability testing theory, a general reliability testing procedure and sample size requirements.

Method A strict protocol was developed, consisting of definitions of the specific reliability measures used, the statistical methods used to assess these reliability coefficients, the development of statistical programmes to make inter-centre reliability comparisons, criteria for good reliability, and a general format for the reliability analysis.

Conclusion The reliability analyses are based on classical test theory. Reliability measures used are Cronbach's α, Cohen's κ and the intraclass correlation coefficient. Intersite comparisons were extended with a comparison of the standard error of measurement. Criteria for good reliability may need to be adapted for this type of study. The consequences of low reliability, and reliability differing between sites, must be considered before pooling data.

Declaration of interest No conflict of interest. Funding detailed in Acknowledgements.

The development and translation of assessment instruments in psychiatry is a lengthy, multi-phase and complex process, which becomes increasingly important in a uniting but still multilingual Europe. A lower or questionable quality of any of the developmental phases makes the interpretation and publication of research based on these instruments a difficult if not impossible task. The major aim of the European Commission BIOMED funded multi-site study called EPSILON (European Psychiatric Services: Inputs Linked to Outcome and Needs) was to produce standardised versions in several European languages of instruments measuring five key concepts in mental health service research (Becker *et al*, 1999): (a) the Camberwell Assessment of Need (CAN), measuring the needs for care; (b) the Involvement Evaluation Questionnaire (IEQ), assessing family or caregiving burden; (c) the Verona Service Satisfaction Scale (VSSS), measuring satisfaction with services; (d) the Lancashire Quality of Life Profile (LQoLP), assessing the quality of life of the patient; and (e) the Client Socio-Demographic and Service Receipt Inventory (CSSRI). This paper describes the background, rationale and design of the reliability study. Its main focus will be on the importance of reliable instruments, theoretical considerations of reliability testing and the general reliability testing procedure of the study.

BACKGROUND

Throughout Europe there is a trend away from hospital-based services towards a variety of locally based community care services for people with severe mental health problems. These community-based services are organisationally more complex, and potentially more demanding on the families of the patients and the local communities. However, they are more likely than hospital-based services to successfully target services to the needs of the most disabled patients, and, as a consequence, are more likely to produce better outcomes at lower treatment and social costs.

Because of its organisational complexity compared with hospital services, community care is more difficult to evaluate. For a proper evaluation of these newer forms of service, a multi-dimensional approach is required (Knudsen & Thornicroft, 1996), which, in addition to well-known patient characteristics like psychopathology and social functioning, should also focus on concepts like needs for care, satisfaction with services, quality of life and family or caregiving burden. To measure these concepts, several instruments have been developed in Europe during the past decade (Schene, 1994; Tansella, 1997). After scientific work on the validity and reliability of the original versions, these instruments were translated into several languages. However, as a consequence of the urgent need to measure the process and the outcome of community care, in most cases the psychometric qualities of these translations (cultural validity and reliability) were not adequately tested.

AIMS AND METHOD OF THE EPSILON STUDY

In an attempt to attack this scientific omission, the aim of this study was to produce standardised versions in five European languages (Danish, Dutch, English, Italian, Spanish) of the five instruments mentioned. These instruments have been developed by different research groups in the past 5–10 years. Because of the quality of the developmental processes, the reliability and validity of each of the original instruments for its particular cultural setting was considered to be good (for the development of each instrument, see the separate reliability papers in this supplement: Chisholm *et al*, 2000; McCrone *et al*, 2000; Ruggeri *et al*, 2000; van Wijngaarden *et al*, 2000; Gaite *et al*, 2000). So the study was focused on generalising the validity and reliability of these instruments over the five European cultural settings. This was done in three steps: (a) translation and back-translation according to World Health Organization standards into the five European languages (Sartorius & Kuyken, 1994); (b) a review of translations by the focus group technique (Vázquez-Barquero & Gaite, 1996) for cultural validity and applicability, and adaptation of the instruments if necessary (for

https://doi.org/10.1192/bjp.177.39.s15 Published online by Cambridge University Press

a more detailed description see Knudsen *et al*, 2000, this supplement); and (c) evaluation of the instruments' reliability (Schene *et al*, 1997). This paper discusses the third step of the procedure.

RELIABILITY

Theoretical considerations

The quality of any measurement instrument (such as an interview or questionnaire) depends on the validity and reliability of the instrument. 'Validity' refers to the extent to which a (test) score matches the actual construct it has to measure, or in other words to the bias or impact of *systematic* errors on test scores. 'Reliability' refers to the extent to which the results of a test can be replicated if the same individuals are tested again under similar circumstances, or, in other words, to the precision and reproducibility or the influence of *unsystematic* (random) errors on the test scores.

Two approaches to reliability can be distinguished: modern test theory (item response theory) and classical test theory. The item response approach makes a comparison of an instrument's performance over different populations possible because, contrary to classical test theory, reliability coefficients in item response theory are not influenced by the population variance. This advantage, however, is diminished by the assumptions concerning the quality of data (e.g. monotonically increasing tracelines, local independence of the items, and – in most cases – dichotomous items), limiting the applicability of an item response approach to those relatively scarce data that fulfil these constraints. In addition, a relatively large number of respondents at each site (200–1000) is needed for an item response approach. The relatively severe constraints on the data, as well as the sample size requirements, made the item response approach not feasible for the EPSILON Study (Donner & Eliasziw, 1987). So it was decided to base the reliability analyses in this study on classical test theory.

In classical test theory, a person's observed score can be expressed as

$$X_i = T_i + E_i$$

(X_i=observed score, T_i=true score and E_i=the error or random, non-systematic part of the score). In psychology and psychiatry, 'gold standards' are lacking, and so a person's true score is defined as his/her theoretical average score over an infinite number of administrations of the same test. Given the assumptions of classical test theory, the variability in the observed scores in a group of respondents (σ_X^2) is composed of the variability in their 'true' scores (σ_T^2) and an error component (σ_E^2). The reliability of a test (ρ_{xx}) is defined as the ratio between true score variance and observed score variance (σ_T^2/σ_X^2). Test reliability defined in the 'classical' way therefore depends to a large extent on the true score variance of the population in which the test was originally developed (since $\sigma_X^2=\sigma_T^2+\sigma_E^2$). If the test is used in another population with a different true score variance (for instance, it might have a lower variance because this population is more homogeneous with respect to the construct under study) the reliability will become lower. For example, in a sample where the error component (σ_E^2) is 0.10 and the true score variance (σ_T^2) is 40, reliability will be 0.80. In another sample with the same error component of 10 but a true score variance of 20, reliability will be 0.66. This 'population' dependence of the reliability coefficient makes comparisons between populations tricky. Differences in reliability between two populations can be caused by differences in precision of the instrument between the populations under study (σ_E^2), or by differences in true score variance of the populations under study (σ_T^2).

One way of handling this problem is the use of the standard error of the mean $(\text{s.e.})_m$, which equals the error component of variance (σ_E). The $(\text{s.e.})_m$, unlike a reliability coefficient, is independent of the true score variance ($(\text{s.e.})_m = (\text{s.d.})_x\sqrt{(1-\rho_{xx})}$, where $(\text{s.d.})_x$ is the standard deviation between subjects). The $(\text{s.e.})_m$ can be interpreted in two ways. First, it can be used to indicate limits within which the observed score would be expected to lie. For example, if the true score were 10, and the $(\text{s.e.})_m$ were 5, for 68% of the time one would expect the observed score to lie in the range 5–15. Second, it indicates the difference to be expected on retesting or between two raters. For example, if the first observed score were 10, the second observed score would be expected to lie in the range 2.9–17.1 ($10 \pm \sqrt{2} \times 5$) for 68% of the time. The $(\text{s.e.})_m$ is therefore particularly useful when assessing the precision of an instrument in absolute terms, in relation to an individual measurement.

Importance of reliable instruments

The reliability of instruments is of importance for at least two reasons.

(a) Low reliability automatically implies low validity. The reliability of a measure is defined as the squared correlation between the observed score X and the true score T ($\rho_{xx}=r_{XT}^2$), a correlation which in case of perfect reliability equals 1. The validity of a measure is defined as the correlation between the true score, T, and the construct one wants to measure, Y. If the validity is perfect, the true score is identical to the actual construct ($T=Y$). Differences between observed scores and Y are only caused by random errors (and hence $r_{XY}=r_{XT}=\sqrt{\rho_{xx}}$). In this case the validity coefficient equals the square root of the reliability coefficient. If validity is not perfect, the value of the validity coefficient will be lower than the square root of reliability; so the reliability coefficient sets the upper limit for the validity coefficient ($r_{XT} \leqslant \sqrt{\rho_{xx}}$).

(b) Unreliability masks the true relationship between constructs under study. If the error components of the observed scores are uncorrelated, the maximal theoretical possible correlation between two unreliable measures is the square root of the product of their respective reliabilities: $r_{X_1X_2} \leqslant \sqrt{(r_{X1X2}r_{X2X2})}$. So research for relationships between different constructs is seriously hampered by unreliable operationalisations of these constructs.

Reliability assessment procedures in the EPSILON Study

In this study three different reliability measures are used, depending on the nature of the instruments involved and the way they are administered (interviews v. questionnaires): (a) Cronbach's α for scales and sub-scales consisting of more than one item; (b) Cohen's κ to estimate the interrater reliability and test–retest reliability of single items; and (c) the intraclass correlation coefficient (ICC) to estimate the interrater reliability and test–retest reliability of scales and sub-scales.

Cronbach's α

If a particular construct is measured by means of a scale consisting of two or more items, measures of internal consistency can be used to estimate the reliability of the

https://doi.org/10.1192/bjp.177.39.s15 Published online by Cambridge University Press

scale. A simple measure of internal consistency is the split-half reliability of a scale, obtained by randomly dividing the scale into two sub-scales and calculating the correlation between those two sub-scales. The Cronbach's α statistic can be considered as the average of all possible split-half reliabilities of a scale. It is sometimes referred to as the internal consistency coefficient (Streiner & Norman, 1995). However one should take into account that α is a function not only of the mean inter-item correlation (a real measure of the internal consistency) but also of the number of items of the scale; hence an increase in α does not automatically mean an increase in the internal consistency. Therefore α can more properly be interpreted as the lower limit of the proportion of variance in the test scores explained by common factors underlying item performance (Crocker & Algina, 1986), such that the lower limit of the reliability – the 'true' reliability – is at least as high as α (Dunn, 1989).

The value of α may be expected to substantially underestimate the reliability if different items measure different quantities (Shrout, 1998); as, for example, in the CAN, where differences between needs in different areas reduce the value of α but do not necessarily imply poor reliability. On the other hand, the errors in individual items in the same scale at the same time may well be positively correlated, which will tend to inflate α relative to the reliability.

Interrater reliability: Cohen's κ and intraclass correlation coefficients

Compared with self-report data, interview data have an additional source of variance that may account for lack of consistency: the interviewer. Although one would prefer an interview, when administered by two different interviewers to the same patient, to produce approximately the same scores – under the assumption that the patient has not changed over time – this is not always the case. Standardisation and structuring of the interview, combined with a thorough training, should in practice diminish the influence of any idiosyncratic characteristics of the interviewers.

The generalisability of the interview scores over interviewers can be estimated by computing a measure of interrater reliability which quantifies the extent to which the information obtained by a specific interviewer can be generalised to other (potential) interviewers. Cohen's κ coefficient is used for categorical data in this study (for variables with more than two categories, a weighted version of the κ coefficient can be used), and ICC for data with at least an ordinal level of measurement.

Strictly speaking, κ is a measure of agreement, not a reliability coefficient, since it is not defined as a ratio of true score variance to observed score variance. κ is defined as $(P_o - P_e)/(1 - P_e)$ where P_o is the observed agreement and P_e is the chance agreement: a value of 0 means that the observed agreement is exactly what could be expected by chance, while a value of 1 indicates perfect agreement.

The ICC is computed as the ratio of between-patient variance to total variance, which is the sum of between-patient variance and error variance (Streiner & Norman, 1995). If systematic bias is present (for example, if one rater systematically reports higher scores than the other), then this is reflected in the ICC.

Test–retest reliability: intraclass correlation coefficient and Cohen's κ

The test–retest reliability coefficient, sometimes called the stability coefficient, tests the assumption that when a characteristic is measured twice, both measures must lead to comparable results. However, test–retest reliability is only a valid indicator of the reliability of an instrument if the characteristic under study has not changed in the interval between testing and retesting. This means either a relatively stable characteristic (like intelligence, personality, socioeconomic status) or a short time interval. A short time interval between test administrations, however, may produce biased (inflated) reliability coefficients, due to the effect of memory.

Crocker & Algina (1986) ask two questions with regard to the interpretation of a stability coefficient as a measure of reliability. First, does a low value of the stability coefficient imply that the test is unreliable or that the construct itself has changed over time? Second, to what extent is an examinee's behaviour or perception of the situation altered by the test administration? In the EPSILON Study we are dealing with relatively stable constructs, so low stability will indicate low reliability. However, some effect of the test administration on a patient's behaviour and/or perception cannot be ruled out. For this reason, the value of the stability coefficient must be considered as a lower limit for the test–retest reliability.

As was the case with interrater reliability, the kind of test–retest statistics used in this study depends on the nature of the instruments. In the case of items containing categorical data (weighted), κ is used. In the case of instruments containing ordinal scales and sub-scales, the ICC statistic is used.

Interviewer characteristics may cause systematic differences between test and retest interview scores. Although reliability, strictly speaking, only refers to unsystematic differences, we believe that the interviewer-related systematic differences should also be taken into account when evaluating the test–retest reliability of the instruments. For this reason we do not use statistics insensitive to systematic change, like rank order correlations, but κ and ICC.

Reliability analysis: design and procedure

Study sites

For this study, researchers from five centres geographically and culturally spread across the European Union (Amsterdam, Copenhagen, London, Santander and Verona) joined forces. All had experience in health services research, mental health epidemiology, and the development and cross-cultural adaptation of research instruments, and had access to mental health services providing care for local catchment areas.

Sample

The following criteria were applied in all centres.

Inclusion criteria: aged between 18 and 65, inclusive, with an ICD–10 F20 diagnosis (schizophrenia), in contact with mental health services during the 3-month period preceding the start of the study.

Exclusion criteria: currently residing in prison, secure residential services or hostels for long-term patients; co-existing learning disability (mental retardation); primary dementia or other severe organic disorder; and extended in-patient treatment episodes longer than one year. These criteria were laid down in order to avoid bias between sites due to variation in the population of patients in long-term institutional care, and to concentrate on those in current 'active' care by specialist mental health teams.

First, administrative prevalence samples of people with ICD–10 diagnosis of any of

https://doi.org/10.1192/bjp.177.39.s15 Published online by Cambridge University Press

Table 1 Reliability testing for each instrument

Instrument	Score distribution: mean & s.d. tests	Internal consistency		Test–retest reliability[1]				Systematic change: paired *t*-test
		α	α test	κ	κ test	ICC	ICC test	
CAN								
items				*				
sumscore	*	*	*			*	*	*
LQoLP								
scales	*	*	*			*	*	*
sumscore	*	*	*			*	*	*
IEQ								
scales	*	*	*			*	*	*
sumscore	*	*	*			*	*	*
VSSS								
items				*	*			
scales	*	*	*			*	*	*
sumscore	*	*	*			*	*	*

1. Also interrater reliability for CAN.
CAN, Camberwell Assessment of Need; LQoLP, Lancashire Quality of Life Profile; IEQ, Involvement Evaluation Questionnaire; VSSS, Verona Service Satisfaction Scale.

F20 to F25 in contact with mental health services were identified either from psychiatric case registers (Copenhagen and Verona) or from the case-loads of local specialist mental health services (in-patient, out-patient and community). Second, cases identified were diagnosed using the item group checklist (IGC) of the Schedule for Clinical Assessment in Neuropsychiatry (SCAN) (World Health Organization, 1992). Only patients with an ICD–10 F20 (schizophrenia) research diagnosis were included in the study. The numbers of patients varied from 52 to 107 between sites, with a total of 404.

For test–retest reliability a randomly selected subsample was tested twice within an interval of 1–2 weeks. The sample sizes differed between sites, ranging from 21 to 77 for the IEQ and from 46 to 81 for the LQoLP. We refer the reader to the separate reliability papers in this supplement for more detailed information (Chisholm *et al*, 2000; Gaite *et al*, 2000; McCrone *et al*, 2000; Ruggeri *et al*, 2000; van Wijngaarden *et al*, 2000).

Core study instruments

The assessment of needs was made using the Camberwell Assessment of Need (CAN) (Phelan *et al*, 1995), which is an interviewer-administered instrument which comprises 22 individual domains of need. The Involvement Evaluation Questionnaire (IEQ) (Schene & van Wijngaarden, 1992) is an 81-item instrument which measures the consequences of psychiatric disorders for relatives of the patient. Caregiving consequences are summarised in four scales: tension, worrying, urging and supervision. The Verona Service Satisfaction Scale (VSSS) (Ruggeri & Dall'Agnola, 1993) is a self-administered instrument comprising seven domains: global satisfaction, skill and behaviour, information, access, efficacy, intervention and relatives' support. The Lancashire Quality of Life Profile (LQoLP) (Oliver, 1991; Oliver *et al*, 1997) is an interview which assesses both objective and subjective quality of life on nine dimensions: work/education, leisure/participation, religion, finances, living situation, legal and safety, family relations, social relations and health. The CSSRI–EU, an adaptation of the Client Service Receipt Inventory (CSRI) (Beecham & Knapp, 1992), is an interview in which socio-demographic data, accommodation, employment, income and all health, social, education and criminal justice services received by a patient during the preceding 6-months are recorded. It allows costing of services received after weighting with unit cost data (for more details about the instruments, see Table 1 in Becker *et al*, 2000).

Reliability protocol

To compare the results from the reliability analyses for the different instruments, a strict protocol was developed (Schene *et al*, 1997) to ensure that all centres used the same procedure and options, and the same software, to test the reliability of instruments and to compare the reliability results of the different centres. The protocol covered the following aspects: definition of the specific reliability measures used; description of the statistical methods to assess these reliability coefficients; development of statistical programmes to make inter-centre reliability comparisons; criteria for good reliability; criteria for pooling v. not pooling data; and the general format for the reliability analysis. In Table 1 the reliability estimates used are presented for all instruments. The justifications for these estimates for each instrument are given in the separate papers in this supplement.

Statistics

Reliability estimates Cronbach's α was computed for each site, using the SPSS reliability module (SPSS 7.5 or higher). ICCs were computed using the SPSS general linear model variance components option with maximum likelihood estimation in SPSS. Patients were entered as random effects, and in case of pooled estimates, the centre was entered as fixed effects. Variance estimates were transformed into ICC estimates with corresponding standard errors using an Excel spreadsheet, inputting the between-patient and error components of variance and their variance–covariance

https://doi.org/10.1192/bjp.177.39.s15 Published online by Cambridge University Press

matrix, the latter being used to obtain standard errors based on the delta technique (Dunn, 1989). Unweighted κ estimates were computed using the SPSS module 'crosstabs', weighted κ using STATA version 5.0 (Statacorp, 1997). The standard error of measurement for a (sub-)scale is computed by substituting Cronbach's α for ρ_{xx} in the formula for the $(\text{s.e.})_m$ given earlier (for α) or directly from the error component of variance (for ICCs).

Inter-site comparisons Tests for differences in α values between sites were performed using the Amsterdam α-testing program ALPHA.EXE (Wouters, 1998, based on Feldt *et al*, 1987). Homogeneity of variance between sites was tested with Levene's statistic. For all scales and subscales, Fisher's Z transformation was applied to ICCs to enable approximate comparisons to be made between sites (Donner & Bull, 1983). Differences between sites were tested for significance by the method of weighting (Armitage & Berry, 1994) before transforming back to the ICC scale. The standard error of measurement was obtained from the 'error' component of variance.

Finally, a paired *t*-test on test–retest data was carried out in order to assess systematic changes from time 1 to time 2. For the separate items of the CAN, test–retest reliability and interrater reliability for each site were computed as pooled κ coefficients. For the separate items of the VSSS, weighted κ values were computed by site and summarised into bands.

Reliability criteria

For a psychological test, standards used for good reliability are often $\alpha \geqslant 0.80$, ICC $\geqslant 0.90$ and $\kappa \geqslant 0.70$. The instruments in this study, however, are not psychological tests, like (for instance) a verbal intelligence test. The constructs they cover are more diffuse than in psychological tests and the boundaries with other constructs (such as unmet needs and quality of life) are less clear. As a consequence, the items constituting these (sub-)scales are more diverse and less closely related than would be the case in a strict, well-defined one-dimensional (sub-)scale. Taking these points into consideration, applying the 'psychological test' standards for good reliability to our instruments seems somewhat unrealistic. Landis & Koch (1977) give some benchmarks for reliability, with 0.81–1.0 termed 'almost perfect', 0.61–8.0 'substantial' and 0.41–0.60 'moderate'. Shrout (1998) suggests revision of these descriptions so that, for example, 0.81–1.0 would be 'substantial' and 0.61–0.80 would be 'moderate'. However, taking account of the special nature of the data in this study, one can consider 0.5 to 0.7 as 'moderate', and 0.7 and over as 'substantial', and these descriptions have informed the discussion of the adequacy of the coefficients.

Pooled v. separate analysis

In a multi-site study such as this, there are many reasons why one might wish to combine data from the different sites: to summarise the reliability analyses, to identify comparable patients in different sites, and to obtain a larger sample for regression analyses. Whether combining data is reasonable depends on the aim of the analysis and on the results of the reliability analysis for each site.

A first aim is to summarise the level of reliability in the study as a whole. Computing a pooled estimate of a reliability coefficient is reasonable if the site-specific coefficients do not differ significantly. Otherwise a pooled estimate would obscure the variations – but, subject to this proviso, it might nevertheless provide a useful summary.

A second aim is to make comparisons between patients from different sites with the same scale scores: for example, in order to compare outcomes between sites adjusted for differences in symptom severity. This requires scale scores for symptom severity to have the same meaning in different sites. Unfortunately the reliability analysis is unable to tell us whether this is the case. Even with perfect reliability, site A might consistently rate the same actual severity higher than site B; yet this might not be apparent from the data if the mean severity was lower in site A.

A third aim of pooling the samples is to have a larger sample on which to conduct correlation or regression analyses. The possibility discussed above (that sites may differ systematically) makes it desirable that these analyses should adjust for site. Differences in reliability are also important in this case. Lack of reliability in outcome variables will decrease precision, and where this differs between sites, weighting might be necessary. For explanatory variables, there is the more serious problem of bias due to their unreliability, which again might differ between sites. These 'untoward effects' of inefficiency and bias are vanishingly small when reliability is moderate (Shrout, 1998), but one would nevertheless wish to ensure that apparent differences between sites were real, and not just due to these effects. A possible solution for the bias problem is to use 'errors in variables' regression, which can adjust for the effects of differing reliabilities at each site. Analyses should, strictly speaking, be carried out on the type of patients for whom reliability has been established. In the present study, the reliability study was nested within the large substantive study, and the inclusion criteria were similar across sites, so there should be no major problem here.

Analysis scheme

For all instruments the following analysis scheme is followed: assess the site-specific reliability estimates (α, ICC, $(\text{s.e.})_m$); test for inter-site differences in reliability estimates; test for inter-site differences in score distribution (mean and variance) (ANOVA, and Levene test).

In addition to the site-specific analyses, pooled reliability estimates are made. Where all estimates are high (say, above 0.9), then small differences in reliability between sites may be statistically significant, yet relatively unimportant in practical terms. However, where reliability is generally lower, or lower for one or more sites, differences in reliability between sites imply that pooled estimates should be treated with great caution. In such cases, it is necessary to extend the inter-site comparisons with a consideration of the site $(\text{s.e.})_m$ values, differences in underlying score distributions, and possible reasons for differences: for example, in the way in which the instrument was applied. Furthermore, any imprecision and bias due to such differences would also need to be taken into account in the analysis of pooled data, in the ways mentioned above.

For the CSSRI a different approach was chosen, because the CSSRI–EU is a new instrument developed for use in a European setting. Since it is an inventory of socio-economic indicators and service variables rather than a multi-item rating scale, the focus is on achieving validity rather than formal reliability (for more details see Chisholm *et al*, 2000, this supplement).

CONCLUSION

Many technical issues surround the choice of measures of reliability. Such measures

https://doi.org/10.1192/bjp.177.39.s15 Published online by Cambridge University Press

are tools which indicate, among other things, the degree to which associations between variables may be diluted; and poor reliability indicates a problem with an instrument when used to quantify associations. Although good reliability does not necessarily indicate a good instrument, reliability studies are one of the best means available to validate our translated instruments.

AART SCHENE, MD, MAARTEN KOETER, PhD, BOB VAN WIJNGAARDEN, MA, Department of Psychiatry, Academic Medical Centre, Amsterdam, The Netherlands; HELLE CHARLOTTE KNUDSEN, MD, Institute of Preventive Medicine, Copenhagen University Hospital, Denmark; MORVEN LEESE, PhD, Section of Community Psychiatry (PRiSM), Institute of Psychiatry, King's College London, UK; MIRELLA RUGGERI, MD, Department of Medicine and Public Health, University of Verona, Italy; IAN R. WHITE, Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London, UK; JOSÉ LUIS VÁZQUEZ-BARQUERO, FRCPsych, Clinical and Social Psychiatry Research Unit, University of Cantabria, Santander, Spain

Correspondence: Professor Aart H. Schene, Academic Medical Centre, Rm. A3.254, PO Box 22700, 1100 DE Amsterdam, The Netherlands. Tel: +31 20 566 2088; fax: +31 20 697 1971

ACKNOWLEDGEMENTS

The following colleagues contributed to the EPSILON Study. Amsterdam: Dr Maarten Koeter, Karin Meijer, Dr Marcel Monden, Professor Aart Schene, Madelon Sijsenaar, Bob van Wijngaarden; Copenhagen: Dr Helle Charlotte Knudsen, Dr Anni Larsen, Dr Klaus Martiny, Dr Carsten Schou, Dr Birgitte Welcher; London: Professor Thomas Becker, Dr Jennifer Beecham, Liz Brooks, Daniel Chisholm, Gwyn Griffiths, Julie Grove, Professor Martin Knapp, Dr Morven Leese, Paul McCrone, Sarah Padfield, Professor Graham Thornicroft, Ian R. White; Santander: Andrés Arriaga Arrizabalaga, Sara Herrera Castanedo, Dr Luis Gaite, Andrés Herran, Modesto Perez Retuerto, Professor José Luis Vázquez-Barquero, Elena Vázquez-Bourgon; Verona: Dr Francesco Amaddeo, Dr Giulia Bisoffi, Dr Doriana Cristofalo, Dr Rosa Dall'Agnola, Dr Antonio Lasalvia, Dr Mirella Ruggeri, Professor Michele Tansella.

This study was supported by the European Commission BIOMED-2 Programme (Contract BMH4-CT95-1151). 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. In Amsterdam, the EPSILON Study was partly supported by a grant from the Nationaal Fonds Geestelijke Volksgezondheid and a grant from the Netherlands Organization for Scientific Research (940-32-007). In Santander the EPSILON Study was partly supported by the Spanish Institute of Health (FIS) (FIS Exp. No. 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

Armitage, P. & Berry, G. (1994) *Statistical Methods in Medical Research* (3rd edn). Oxford: Blackwell Scientific.

Becker, T., Knapp, M., Knudsen, H. C., *et al* (1999) The EPSILON Study of schizophrenia in five European countries: design and methodology for standardising outcome measures and comparing patterns of care and service costs. *British Journal of Psychiatry*, **175**, 514–521.

—, —, —, *et al* (2000) Aims, outcome measures, study sites and patient sample. EPSILON Study 1. *British Journal of Psychiatry*, **177** (suppl. 39), s1–s7.

Beecham, J. & Knapp, M. (1992) Costing psychiatric interventions. In *Measuring Mental Health Needs* (eds G. Thornicroft, C. R. Brewin & J. Wing), pp. 163–183. London: Gaskell.

Chisholm, D., Knapp, M. R. J., Knudsen, H. C., *et al* (2000) The Client Socio-Demographic and Service Receipt Inventory: development of an instrument for international research. Epsilon Study 5. *British Journal of Psychiatry*, **177** (suppl. 39), s28–s33.

Crocker, L. & Algina, J. (1986) *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart & Winston.

Donner, A. & Bull, S. (1983) Inferences concerning a common intraclass correlation coefficient. *Biometrics*, **39**, 771–775.

— & Eliasziw, M. (1987) Sample size requirements for reliability studies. *Statistics in Medicine*, **6**, 441–448.

Dunn, G. (1989) *Design and Analysis of Reliability Studies*. London: Edward Arnold.

Feldt, L., Woodruff, D. & Salih, F. (1987) Statistical inference for coefficient alpha. *Applied Psychological Measurement*, **11**, 93–103.

Gaite, L., Vázquez-Barquero, J. L., Arriaga Arrizabalaga, A., *et al* (2000) Quality of life in schizophrenia: development, reliability and internal consistency of the Lancashire Quality of Life Profile – European Version. EPSILON Study 8. *British Journal of Psychiatry*, **177** (suppl. 39), s49–s54.

Knudsen, H. C. & Thornicroft, G. (1996) *Mental Health Service Evaluation*. Cambridge: Cambridge University Press.

—, Vázquez-Barquero, J. L., Welcher, B., *et al* (2000) Translation and cross-cultural adaptation of outcome measurements for schizophrenia. EPSILON Study 2. *British Journal of Psychiatry*, **177** (supp. 39), s8–s14.

Landis, J. R. & Koch, G. G. (1977) The measurement of agreement for categorical data. *Biometrics*, **33**, 159–74.

McCrone, P., Leese, M., Thornicroft, G., *et al* (2000) Reliability of the Camberwell Assessment of Need – European Version. EPSILON Study 6. *British Journal of Psychiatry*, **177** (suppl. 39), s34–s40.

Oliver, J. (1991) The social care directive: development of a quality of life profile for use in the community services for the mentally ill. *Social Work & Social Sciences Review*, **3**, 5–45.

—, Huxley, P., Priebe, S. & Kaiser, W. (1997) Measuring the quality of life of severely mentally ill people using the Lancashire Quality of Life Profile. *Social Psychiatry and Psychiatric Epidemiology*, **32**, 76–83.

Phelan, M., Slade, M., Thornicroft, G., *et al* (1995) The Camberwell Assessment of Need: the validity and reliability of an instrument to assess the needs of people with severe mental illness. *British Journal of Psychiatry*, **167**, 589–595.

Ruggeri, M. & Dall'Agnola, R. (1993) The development and use of the Verona Expectations for Care Scale (VECS) and the Verona Service Satisfaction Scale (VSSS) for measuring expectations and satisfaction with community-based psychiatric services in patients, relatives and professionals. *Psychological Medicine*, **23**, 511–523.

—, Lasalvia, A., Dall'Agnola, R., *et al* (2000) Development, internal consistency and reliability of the Verona Service Satisfaction Scale – European Version. EPSILON Study 7. *British Journal of Psychiatry*, **177** (suppl. 39), s41–s48.

Sartorius, N. & Kuyken, W. (1994) Translations of health status instruments. In *Quality of Life Assessments: International Perspectives* (eds J. Orley & W. Kuyken), pp. 19–32. Berlin, Heidelberg: Springer.

Schene, A. H. (1994) *Report of the First International ENMESH Conference: Mental Health Service Evaluation: Developing Reliable Measures*. Amsterdam: Department of Psychiatry, Academic Medical Centre, University of Amsterdam.

— & van Wijngaarden, B. (1992) *The Involvement Evaluation Questionnaire*. Amsterdam: Department of Psychiatry, Academic Medical Centre, University of Amsterdam.

—, Koeter, M. & van Wijngaarden, B. (1997) *Assessing Needs and Cost-Effectiveness of Care for People Severely Disabled by Schizophrenia in the EU: Reliability Protocol*. Amsterdam: Department of Psychiatry, Academic Medical Centre, University of Amsterdam.

Shrout, P. E. (1998) Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, **7**, 301–307.

Statacorp (1997) *Stata Statistical Software Release 5.0*. College Station, TX: Stata Corporation.

Streiner, D. & Norman, G. (1995) *Health Measurement Scales: A Practical Guide to their Development and Use*. Oxford: Oxford University Press.

Tansella, M. (ed.) (1997) *Making Rational Mental Health Services*. Epidemiologia e Psichiatria Sociale, Monograph Supplement 1. Rome: Il Pensiero Scientifico Editore.

van Wijngaarden, B., Schene, A. H., Koeter, M., *et al* (2000) Caregiving in schizophrenia: development, internal consistency and reliability of the Involvement Evaluation Questionnaire – European Version. EPSILON Study 4. *British Journal of Psychiatry*, **177** (suppl. 39), s21–s27.

Vázquez-Barquero, J. L. & Gaite, L. (1996) *Focus Group Interview. Final Protocol*. Santander: Clinical and Social Psychiatry Research Unit, University of Cantabria.

World Health Organization (1992) *Schedules for Clinical Assessment in Neuropsychiatry* (ed.-in-chief J. K. Wing). Geneva: WHO.

Wouters, L. (1998) *Amsterdam Alpha-testing Program ALPHA.EXE*. Amsterdam: Academic Medical Centre.