

## Smallest real difference, a link between reproducibility and responsiveness

H. Beckerman<sup>1</sup>, M.E. Roebroek<sup>2</sup>, G.J. Lankhorst<sup>1</sup>, J.G. Becher<sup>1</sup>, P.D. Bezemer<sup>3</sup> & A.L.M. Verbeek<sup>4</sup>

<sup>1</sup>*Department of Rehabilitation Medicine, University Hospital Vrije Universiteit, Amsterdam (E-mail: h.beckerman@vumc.nl)*; <sup>2</sup>*Institute of Rehabilitation Medicine, Erasmus University, Rotterdam*; <sup>3</sup>*Department of Clinical Epidemiology and Biostatistics, Vrije Universiteit, Amsterdam*; <sup>4</sup>*Department of Epidemiology, University of Nijmegen, The Netherlands*

Accepted in revised form 16 September 2001

### Abstract

The aim of this study is to show the relationship between test–retest reproducibility and responsiveness and to introduce the smallest real difference (SRD) approach, using the sickness impact profile (SIP) in chronic stroke patients as an example. Forty chronic stroke patients were interviewed twice by the same examiner, with a 1-week interval. All patients were interviewed during the qualification period preceding a randomized clinical trial. Test–retest reproducibility has been quantified by the intraclass correlation coefficient (ICC), the standard error of measurement (SEM) and the related smallest real difference (SRD). Responsiveness was defined as the ratio of the clinically relevant change to the SD of the within-stable-subject test–retest differences. The ICC for the total SIP was 0.92, whereas the ICCs for the specified SIP categories varied from 0.63 for the category ‘recreation and pastime’ to 0.88 for the category ‘work’. However, both the SEM and the SRD far more capture the essence of the reproducibility of a measurement instrument. For instance, a total SIP score of an individual patient of 28.3% (which is taken as an example, being the mean score in the study population) should decrease by at least 9.26% or approximately 13 items, before any improvement beyond reproducibility noise can be detected. The responsiveness to change of a health status measurement instrument is closely related to its test–retest reproducibility. This relationship becomes more evident when the SEM and the SRD are used to quantify reproducibility, than when ICC or other correlation coefficients are used.

**Key words:** Outcome measure, Reliability, Reproducibility, Responsiveness, Sickness impact profile, Stroke

**Abbreviations:** ANOVA – analysis of variance; COTA – Center for Orthopedic Techniques Amsterdam; d – days; GENOVA – generalized analysis of variance; ICC – Intraclass Correlation Coefficient; LBP – low back pain; MI – myocardial infarction; mo – months; SD – standard deviation; SEM – Standard Error of Measurement; SIP – Sickness Impact Profile; SRD – Smallest Real Difference; var – variance

### Introduction

Outcome measures should be stable in stable subjects (reproducibility), but should also be able to detect changes in unstable subjects (responsiveness). Reproducibility, which is also known as the test–retest reliability, and responsiveness are

closely related methodological properties of an outcome measure. Reproducibility as well as responsiveness have to do with repeated measures of the same individual. Better reproducibility implies better precision of single measurements, which is a prerequisite for better tracking of changes in measurements in research or clinical practice [1].

Reproducibility is defined as the ability to measure attributes in a consistent manner when administered on several occasions to stable subjects [2, 3]. Reproducibility is not a singular quality of a measurement instrument, but is dependent on the design of the reproducibility study (e.g. the timing of the measurements and the number of facets, i.e. sources of variance) and the study population. Responsiveness is the ability of a scale to detect clinically relevant changes over time [3]. There are several techniques for the quantification of responsiveness [4]. According to Guyatt et al. [3] responsiveness is the ratio of the clinically relevant change to the SD of the within-stable-subject test–retest differences. Thus, the test–retest reproducibility of an instrument has direct implications for its responsiveness.

In clinical practice as well as in clinical trials, information is needed to answer the question of how much difference is needed to detect a real change, considering chance variation or measurement error [5]. To address this important issue, a test–retest reproducibility study in chronic stroke patients was performed during the qualification period of a randomized clinical trial, using the sickness impact profile (SIP) as an example. The objective of this study was to investigate the relationship between reproducibility and responsiveness, and to introduce the smallest real difference (SRD) approach.

## Methods

### *Study population*

The study population consisted of candidates for a randomized clinical trial concerning the efficacy of percutaneous radiofrequency thermocoagulation of the tibial nerve and an ankle foot orthosis on walking ability [6]. Included in the study were patients aged between 18 and 75 years, who at least 4 months previously had suffered an ischemic or hemorrhagic stroke of a cerebral hemisphere resulting in hemiplegia, and were experiencing walking problems caused by a spastic equinus or an equinovarus position of the foot. Participants were also required to have adequate communication and cognitive abilities. The research protocol was approved by the medical ethics committee of the University Hospital Vrije Universiteit.

### *Sickness impact profile*

The SIP has been developed to provide an appropriate and responsive measure of health status, for use in assessing the effects of health care services [7–9]. Published data on the test–retest reproducibility of the SIP are often extracted from the original publications of McMahon and Dawborn [10], Hazard Munro et al. [11] and Hyde [12]. However, Bergner et al. [7, 13, 14] and Bruin et al. [15] used earlier versions of the SIP, containing 312, 236 and 189 items respectively. Few authors have used the final version of the SIP to investigate reproducibility [16–22]. This final version includes a physical and a psychosocial dimension, 12 categories, and 136 items [14, 15]. The results of these studies have shown correlation coefficients ranging from 0.20 to 0.95 [16–22]. With the exception of the study carried out by Jacobs et al. [20], none of the reproducibility studies discussed the consequences of the test–retest results, i.e. incomplete reproducibility (coefficients <1), in relation to the responsiveness of the SIP [16–22].

In our study we used the validated Dutch version of the SIP. This version has the same weight factors as the original American version [23, 24]. The SIP covers 12 different categories of disabilities, based on 136 yes/no answers, and a weighted score is applied to each answer. Furthermore, a physical dimension score (based on three categories) and a psychosocial dimension score (based on four categories) can be calculated. The items refer to various aspects of health-related dysfunctioning in the patient on the day of the interview. Potential scores range from 0 to 100%, with higher percentages representing greater ‘sickness impact’, and thus poorer health [14, 23, 24].

### *Design of the test–retest reproducibility study*

During the qualification period preceding the clinical trial, patients were interviewed on two separate occasions with a 1-week interval. The first measurement took place prior to a diagnostic block of the tibial nerve with 0.5% marcaine [25]. The effect of this local anesthetic disappears after a few hours. The second measurement took place 1 week later, prior to randomization. All patients were interviewed by the same person. This implied a one-facet study design, with ‘week’ (i.e. mea-

surement) as the only source of variance [1, 26]. No clinically relevant changes in the health status of the patients were expected during the 1-week interval between the two interviews. Furthermore, because of the diversity of the many statements and the time required to complete the SIP, it was expected that patients would not remember their first responses.

#### *Data-analysis*

The total SIP scores, as well as the category and dimension scores, are considered as interval scales [27]. Test–retest reproducibility was assessed by calculating three different numerical indexes [1, 28]. All these reproducibility indexes are based on analyses of variance components. The total variance can be sub-divided into (1) variance between subjects and (2) variance within subjects. Using this one-facet study design, the within-subject variance could be sub-divided into two components: ‘week’ (i.e. week 1 and week 2) and ‘residual’, respectively. This last component includes the two-way interaction for ‘subject  $\times$  week’ as well as a certain amount of residual error resulting from systematic and unsystematic error sources that are unknown [26]. The following numerical indexes were calculated:

(1) Intraclass correlation coefficient (ICC): The ICC is generally accepted in the medical literature as the preferred method of quantifying reproducibility [3, 29]. The ICC is calculated as the ratio of the variance between subjects (i.e. the variance of interest) and the total variance.

(2) Standard error of measurement (SEM): Quantifying the test–retest reproducibility of an assessment involves calculating the variability in measurements of the same individual, i.e. the variance within subjects, which is referred to as error variance. The SEM is the square root of the within-subject variance (i.e. the square root of the total variance, excluding the variance between subjects) [1, 2, 30], and is expressed in the same dimension as the measurement (note: the standard error of measurement (SEM) is not the same as the standard error of the mean, which is also abbreviated as SEM).

(3) Smallest real difference (SRD): Measurement error makes the observed value of a measure differ

from the true value. Assuming that measurement errors are distributed normally, an interval or error band can be calculated around one measurement, expressing the uncertainty of the observed score, as induced by error [1, 2]. At the 95% confidence level, this interval is equal to  $\pm 1.96 \times \text{SEM}$ . To calculate an error band around the difference between two measurements, we assume that the measurement error of both measurements are independent of each other. This interval or error band expresses the uncertainty of the difference between the two observed scores [1, 28].

When this interval contains the value 0, the difference between the two measurements could be induced by error alone. The difference between both measurements should be at least  $1.96 \times \sqrt{2} \times \text{SEM}$  to indicate 95% confidence of a real difference between the true scores (1.96 because of the 95% confidence,  $\sqrt{2}$  because of the difference of two variances). We propose to call the index  $1.96 \times \sqrt{2} \times \text{SEM}$  the ‘SRD’. In other words, the SRD is the smallest measurement change, that can be interpreted as a real difference, i.e. beyond zero [31, 32].

To calculate the ICC, the SEM and the SRD, analysis of variance (ANOVA) for random effects was performed, using the PC version of the GENOVA programme, developed by Crick and Brennan [33, 34].

## **Results**

### *Study population*

The study population consisted of 40 chronic stroke patients referred to the outpatient Department of Rehabilitation Medicine of the University Hospital of the Vrije Universiteit. Ten women and 30 men, with a median age of 58 years (range: 26–72 years), and with a median interval between the cerebrovascular accident and the first measurement of 45.5 months (range: 5–185 months) gave their informed consent to participate in the study. Six patients had experienced a hemorrhagic stroke and 34 patients had experienced an ischemic stroke. Half of the patients had left-sided hemisphere lesions, and the other half had right-sided hemisphere lesions. Almost 70% of the patients

also had some type of comorbidity (musculoskeletal system, tractus respiratorius, tractus digestivus, peripheral circulatory system, cardiovascular system, diabetes mellitus, etc.).

### *SIP scores*

Descriptive statistics, including the mean and SD for the total SIP scores, for the physical and psychosocial dimensions and for each category, are presented in Table 1. The total SIP score varied from 8.5 to 51.1%, the physical dimension score varied from 6.7 to 59.2%, and the psychosocial dimension score varied from 0.0 to 59.6% (first interview). Patients stated that their stroke had great impact on their walking ability, and the mean ambulation score is quite high. This is not surprising, because to be included in the clinical trial, one of the criteria was that patients experienced problems with walking. Recreation and pastime, work, and household management were also greatly affected.

### *Test-retest reproducibility*

With ANOVA for random effects, the between-subject and within-subject variance components were computed. These variance components are listed in Table 2, together with the three repro-

ducibility indexes. The SRD is expressed both as a percentage and as an estimated number of SIP items (last column of Table 2). Although each SIP item has a different weight, we estimated the number of SIP items that approximately corresponds with the SRD%, by calculating the mean item-weight in percentages for each SIP category (i.e. 100%/total number of items in the category), and dividing this percentage by the SRD%.

The results of the current study indicate that, with the exception of the categories emotion, sleep and rest, and household management, the mean differences between the first and the second measurement are small (last column of Table 1). The ICCs for the SIP categories and dimensions are ranging from 0.922 for the total SIP score to 0.633 for recreation and pastime (Table 2).

However, the SEMs and the SRDs give a better view of the test-retest reproducibility of the SIP. The SRDs, expressed in the same dimension as the SIP, are large (Figure 1), especially when taking into account the baseline values of the SIP (see Table 1). For instance, the score of a patient with a total SIP score of 28.3% (the same as the group mean score in this study) should decrease by at least 9.26% (or an approximate reduction of 13 'positive' SIP items) before any real improvement is detected (Figure 2). Smaller changes should be interpreted as measurement error or 'reproduc-

**Table 1.** Descriptive statistics of the SIP scores (%) on two separate occasions with 1-week interval

Areas of disability	First measurement		Second measurement 1-week later		Difference week 1-2	
	Mean	SD	Mean	SD	Mean	SD
Physical dimension	29.8	13.4	30.9	15.2	-1.1	6.4
Ambulation	43.3	12.8	45.2	14.3	-1.9	7.6
Body care and movement	24.6	14.4	26.1	16.3	-1.4	7.6
Mobility	28.6	20.9	27.8	22.5	0.8	12.7
Psychosocial dimension	24.1	14.5	22.6	16.9	1.5	8.7
Social interaction	19.4	13.0	19.3	15.7	0.1	10.3
Emotion	23.3	21.8	19.0	22.1	4.2	12.4
Alertness	30.3	27.0	27.5	29.4	2.8	20.5
Communication	27.6	22.6	27.1	22.8	0.4	15.5
Independent categories						
Sleep and rest	21.6	18.2	16.8	14.7	4.9	8.5
Household management	44.3	21.9	48.5	22.0	-4.3	13.8
Recreation and pastime	44.6	19.8	42.7	18.1	1.9	16.3
Eating	11.1	7.6	10.4	8.2	0.8	6.5
Work	43.2	32.9	42.3	32.9	0.9	16.3
Total SIP	28.3	11.0	28.0	12.9	0.3	4.7

**Table 2.** Test–retest reproducibility results of the SIP (N = 40)

Areas of disability <sup>a</sup>	Between-subject variance	Within-subject var		ICC	SEM (%)	SRD (%)	SRD number of items <sup>b</sup>
		Week	Residue				
Physical dimension (45)	184.08	0.08	20.49	0.899	4.54	12.57	6
Ambulation (12)	156.37	1.08	28.54	0.841	5.44	15.08	2
Body care and movement (23)	207.44	0.31	28.93	0.876	5.41	14.99	4
Mobility (10)	390.49	0.00	80.41	0.829	8.97	24.86	3
Psychosocial dimension (48)	210.78	0.21	37.64	0.848	6.15	17.05	9
Social interaction (20)	155.28	0.00	52.76	0.746	7.26	20.13	4
Emotion (9)	405.10	7.04	77.03	0.828	9.17	25.42	3
Alertness (10)	586.35	0.00	211.09	0.735	14.53	40.27	4
Communication (9)	396.58	0.00	120.00	0.768	10.95	30.36	3
Independent categories (43)							
Sleep and rest (7)	238.40	10.93	36.51	0.834	6.89	19.09	2
Household management (10)	387.61	6.83	94.84	0.792	10.08	27.95	3
Recreation and pastime (8)	228.16	0.00	132.05	0.633	11.49	31.85	3
Eating (9)	41.51	0.00	21.28	0.661	4.61	12.79	2
Work (9)	950.47	0.00	133.54	0.877	11.56	32.03	–
Total SIP (136)	131.94	0.00	11.16	0.922	3.34	9.26	13

var – variance.

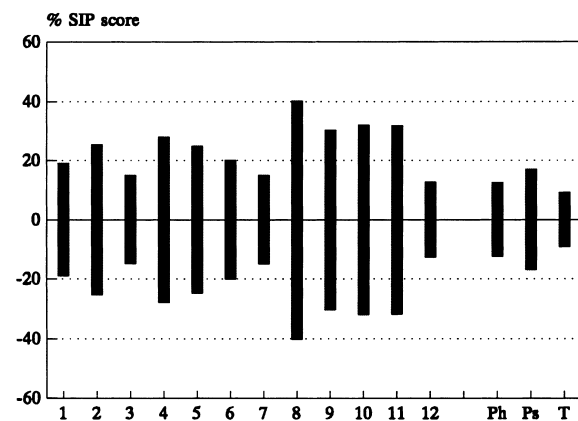
$$ICC = \frac{\text{between-subject variance}}{\text{between-subject variance} + \text{within-subject variance}}$$

$$SEM = \sqrt{\text{within-subject variance}} = \sqrt{(\text{total variance})(1 - ICC)}$$

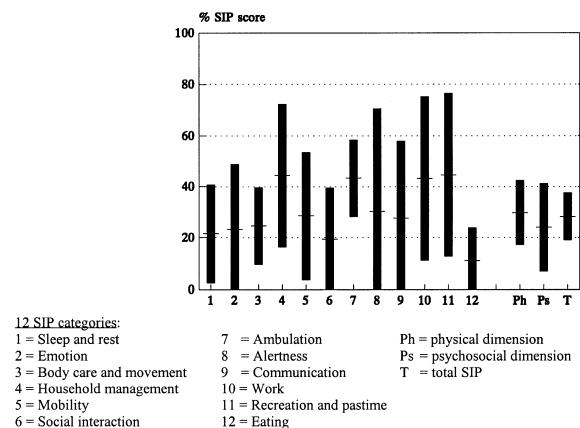
$$SRD = 1.96 \times \sqrt{2} \times SEM.$$

<sup>a</sup>The total number of items in each dimension and category is given between parentheses.

<sup>b</sup>The SRD is expressed as an estimated number of SIP items, by calculating the mean item-weight in percentages for each SIP category (i.e. 100%/total number of items in the category), and dividing this percentage by the SRD%. For example, the mean item-weight of the category ambulation is estimated as 100%/12 items = 8.33% per item; the SRD expressed as an estimated number of SIP items is 15.08%/8.33% = 1.81 or approximately two items.



**Figure 1.** Smallest Real Difference (%) for each SIP category, the physical and psychosocial dimension, and the total SIP. Negative score indicating improvement, positive score indicating deterioration.



**Figure 2.** Smallest Real Difference (%) illustrated around the mean SIP scores (%) on the first measurement; higher scores on the second measurement indicate deterioration, lower scores on the second measurement indicate improvement.

ibility noise'. For ambulation, the score should decrease by 15.08% (consistent with approximately two items) before any real improvement is detected (Figure 2).

## Discussion

### *Relationship between reproducibility and responsiveness*

Reproducibility as well as responsiveness refer to within-subject variability, i.e. repeated measures of the same individual. Indexes that capture this notion of within-subject variability are the SEM and the SRD. The ICC is an accepted method of quantifying reproducibility. The ICC is more appropriate than Pearson's product moment correlation coefficient, because systematic variability is neglected in the latter method [29]. However, using the ICC, the variance between subjects is usually considered as variance of interest, whereas with respect to longitudinal changes the magnitude of within-subject variance over time is relevant [28, 31, 35–37]. Furthermore, reproducibility coefficients, expressed as a dimensionless number between 0 and 1, do not lend themselves to straightforward interpretation. As has been demonstrated in this article, reproducibility coefficients such as the ICC are not appropriate for gaining insight into the methodological quality of instruments measuring change within a subject over time. In our opinion, the SEM and the SRD are better suited for this purpose [1, 38].

Results of clinical trials depend on the choice of the primary outcome measure. As mentioned earlier, responsiveness is the ability of a scale to detect clinically relevant changes over time due to some intervention or change in clinical status, and could be expressed as the ratio of the clinically relevant change to the SD of the within-stable-subject test–retest differences [3]. The test–retest reproducibility study delivered us the denominator of this responsiveness ratio. However, with respect to the numerator 'the clinically relevant change' there is no solution yet. At the moment, there are no standardized methods for defining the magnitude of minimal clinically relevant changes [39]. Nor can our SRD approach answer the question what

is clinically important. There is an essential difference between 'clinically relevant change' and the 'SRD'. The SRD is a clinimetric property of a measurement instrument, whereas 'clinically relevant change' is an arbitrarily chosen amount of change indicating which change clinicians and researchers minimally expect or judge as important, taking the natural course, the strength of an intervention, etc., into account. However, the SRD tells you whether the instrument will be able to measure such a difference. If the minimally clinically important differences we want to measure do not exceed the SRD, there is no doubt that the measurement instrument is not valid for this purpose [40]. Therefore, we propose the following approach: if the numerator of the responsiveness ratio of Guyatt et al. [3] is set equal to the SRD, the responsiveness is by definition 1.96. In this respect, the SRD approach may be helpful for making sample-size calculations in the planning of clinical trials and a critical evaluation of which measure to define as the primary outcome measure [1]. Based on the test–retest results of this study, measuring true changes within a chronic stroke patient seems to be almost impossible with the SIP, especially when taking into account the baseline values. Hence, using the SIP very large treatment effects are needed to detect efficacy. As has been reported in previous studies, there are some doubts about the responsiveness of the SIP in stroke patients. For instance, Schuling et al. [41] found no changes in SIP scores during the first 6 months post-stroke, whereas during the same period the Barthel scores changed significantly. It was concluded that the SIP was not responsive enough to detect the modest improvement in a consecutive cohort of acute stroke patients [41].

### *Improving reproducibility and responsiveness*

Measurement error makes the observed value of a measure differ from the true value. Measurement error not only affects a single measurement, but also affects the measurement of changes (responsiveness). An instrument that shows high variability within stable subjects would be considered to have poor responsiveness. In this respect, re-

producibility is a necessary condition of responsiveness. It should be noted that reproducibility as well as responsiveness indexes are no fixed characteristics of a measurement instrument. Factors associated with the study design (e.g. the time-interval, method of administration, number of observers), the study population (e.g. diagnosis, age, gender, emotional status, cognitive level), therapeutic interventions, etc., might all influence the magnitude of the variance between subjects as well as the error variance. The most valid approaches to improve reproducibility and responsiveness are those which reduce measurement error. The generalizability theory helps to analyse the sources and the relative magnitude of measurement error [2, 26, 28]. The effects of specific strategies to reduce measurement error (e.g. training of the observers, repeated measurements) can also be analysed. In order to reduce measurement error, two or more repeated measurements are preferable. By calculating the average of  $k$  scores, the SEM is reduced by a factor  $1/\sqrt{k}$  [2]. However, for the SIP it is not recommended that a second measurement should take place shortly after the first one, because patients experience the SIP as lengthy and boring. If it is included as part of a series of measurements in a clinical trial the burden on the respondent might also affect the other measurements. An alternative strategy for improving the responsiveness has been demonstrated by Stratford et al. [42], who identified 20 SIP items that proved to be responsive to treatment in low back pain patients. Although, by selecting the most responsive items and by changing the items by adding 'because of your back pain', the generic health status instrument becomes more disease-specific. This set bounds to the usefulness of the instrument and generalization of the scores to patients with different diagnoses.

## Conclusion

The responsiveness of a health status measurement instrument is closely related to its test-retest reproducibility. This relationship becomes more evident when the SEM and in particular the SRD are used to quantify reproducibility, than when ICCs or other correlation coefficients are used.

## Acknowledgement

This study is part of a project that has been supported by a grant from the Netherlands Heart Foundation (project 91.060).

## References

1. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med* 2000; 30: 1–15.
2. Streiner DL, Norman GR. *Health Measurement Scales. A Practical Guide to Their Development and Use*. Oxford: Oxford University Press, 1993.
3. Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. *J Chron Dis* 1987; 40: 171–178.
4. Bennekom CAM van, Jelles F, Lankhorst GJ, Bouter LM. Responsiveness of the rehabilitation activities profile and the Barthel index. *J Clin Epidemiol* 1996; 49: 39–44.
5. Moertel CG, Hanley JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer* 1976; 38: 388–394.
6. Beckerman H, Becher J, Lankhorst GJ, Verbeek ALM. Walking ability of stroke patients: Efficacy of tibial nerve blocking and a polypropylene ankle-foot orthosis. *Arch Phys Med Rehabil* 1996; 77: 1144–1151.
7. Gilson BS, Gilson JS, Bergner M, et al. The sickness impact profile. Development of an outcome measure of health care. *Am J Public Health* 1975; 65: 1304–1310.
8. Bergner M, Bobbitt RA, Kressel S, Pollard WE, Gilson BS, Morris JR. The sickness impact profile: Conceptual formulation and methodology for the development of a health status measure. *Int J Health Serv* 1976; 6: 393–415.
9. Bergner M, Bobbitt RA, Pollard WE, Martin DP, Gilson BS. The sickness impact profile: Validation of a health status measure. *Med Care* 1976; 14: 57–67.
10. McMahon LP, Dawborn JK. Subjective quality of life assessment in hemodialysis patients at different levels of hemoglobin following use of recombinant human erythropoietin. *Am J Nephrol* 1992; 12: 162–169.
11. Hazard Munro B, Creamer AM, Haggerty MR, Cooper FS. Effect of relaxation therapy on post-myocardial infarction patients' rehabilitation. *Nurs Res* 1988; 37: 231–235.
12. Hyde E. Acupressure therapy for morning sickness. A controlled clinical trial. *J Nurse-Midwifery* 1989; 34: 171–178.
13. Pollard WE, Bobbitt RA, Bergner M, Martin DP, Gilson BS. The sickness impact profile: Reliability of a health status measure. *Med Care* 1976; 14: 146–155.
14. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The sickness impact profile: Development and final revision of a health status measure. *Med Care* 1981; 19: 787–805.
15. Bruin AF de, Witte LP de, Stevens F, Diederiks JPM. Sickness impact profile: The state of the art of a generic functional status measure. *Soc Sci Med* 1992; 35: 1003–1014.
16. Deyo RA, Diehl AK. Measuring physical and psychosocial function in patients with low back pain. *Spine* 1983; 8: 635–642.

17. Deyo RA. Comparative validity of the sickness impact profile and shorter scales for functional assessment in low-back pain. *Spine* 1986; 11: 951–954.
18. Jensen MP, Strom SE, Turner JA, Romano JM. Validity of the sickness impact profile Roland scale as a measure of dysfunction in chronic pain patients. *Pain* 1992; 50: 157–162.
19. Sullivan M, Ahlmen M, Archenholtz B, Svensson G. Measuring health in rheumatic disorders by means of a Swedish version of the sickness impact profile. *Scand J Rheumatol* 1986; 15: 193–200.
20. Jacobs HM, Touw-Otten FWMM, Melker RA de. The evaluation of changes in functional health status in patients with abdominal complaints. *J Clin Epidemiol* 1996; 49: 163–171.
21. Damiano AM, Patrick DL, Guzman GI, et al. Measurement of health-related quality of life in patients with amyotrophic lateral sclerosis in clinical trials of new therapies. *Med Care* 1999; 37: 15–26.
22. Visser MC, Koudstaal PJ, Erdmann RA, et al. Measuring quality of life in patients with myocardial infarction or stroke: A feasibility study of four questionnaires in The Netherlands. *J Epidemiol Commun Health* 1995; 49: 513–517.
23. Witte L de, Jacobs H, Horst F van der, Luttik A, Joosten J, Philipsen H. De waarde van de sickness impact profile als maat voor het functioneren van patiënten. *Gezondheid en Samenleving* 1987; 8: 120–127.
24. Jacobs HM, Luttik A, Touw-Otten FWMM, Kastein M, De Melker RA. Measuring impact of sickness in patients with nonspecific abdominal complaints in a Dutch family practice setting. *Med Care* 1992; 30: 244–251.
25. Arendzen JH, Duijn H van, Beckmann MKF, Harlaar J, Vogelaar TW, Prevo AJH. Diagnostic blocks of the tibial nerve in spastic hemiparesis. Effects on clinical, electrophysiological and gait parameters. *Scand J Rehab Med* 1992; 24: 75–81.
26. Shavelson RJ, Webb NM. *Generalizability Theory: A Primer*. Newbury Park: Sage Publications, 1991.
27. Carter WB, Bobbitt RA, Bergner M, Gilson BS. Validation of an interval scaling: The sickness impact profile. *Health Services Res* 1976; 517–528.
28. Roebroek ME, Harlaar J, Lankhorst GJ. The application of the generalizability theory to reliability assessment: An illustration using isometric force measurement. *Phys Ther* 1993; 73: 386–401.
29. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Controlled Clin Trials* 1991; 12: 142S–158S.
30. Stratford PW. Reliability: Consistency or differentiating among subjects? *Phys Ther* 1989; 69: 299–300.
31. Guyatt GH, Kirschner B, Jaeschke R. Measuring health status: What are the necessary measurement properties? *J Clin Epidemiol* 1992; 45: 1341–1345.
32. Pfenning LEMA, Ploeg HM van der, Cohen L, Polman CH. A comparison of responsiveness indices in multiple sclerosis patients. *Qual Life Res* 1999; 8: 481–489.
33. Brennan RL. *Elements of Generalizability Theory*. Iowa city: ACT publications, 1983.
34. Crick JE, Brennan RL. *Manual for GENOVA: A Generalized Analysis of Variance System*. Iowa city: American College Testing Program, 1983.
35. Kramer MS, Feinstein AR. Clinical biostatistics. LIV. The biostatistics of concordance. *Clin Pharmacol Ther* 1981; 29: 111–123.
36. Bartko JJ. On various intraclass correlation reliability coefficients. *Psychol Bull* 1976a; 83: 762–765.
37. Bartko JJ, Carpenter WT. On the methods and theory of reliability. *J Nerv Mental Dis* 1976; 163: 307–317.
38. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: An illustration of appropriate statistical analyses. *Clinic Rehabil* 1998; 12: 187–199.
39. Lachs MS. The more things change.... *J Clin Epidemiol* 1993; 46: 1091–1092.
40. Hébert R, Spiegelhalter DJ, Brayne C. Setting the minimal metrically detectable change on disability rating scales. *Arch Phys Med Rehabil* 1997; 78: 1305–1308.
41. Schuling J, Greidanus J, Meyboom-de Jong B. Measuring functional status of stroke patients with the sickness impact profile. *Disabil Rehabil* 1993; 15: 19–23.
42. Stratford P, Solomon P, Binkley J, Finch E, Gill C. Sensitivity of sickness impact profile items to measure change over time in a low back pain patient group. *Spine* 1993; 18: 1723–1727.

*Address for correspondence:* Ms H. Beckerman, Department of Rehabilitation Medicine, University Hospital Vrije Universiteit, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands  
Phone: +31-20-44-40762; Fax: +31-20-44-40787  
E-mail: h.beckerman@vumc.nl