STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:1061–1076 (DOI: 10.1002/sim.723)

Revision of the ICIDH Severity of Disabilities Scale by data linking and item response theory

S. van Buuren*,† and M. Hopman-Rock

TNO Prevention and Health, P.O. Box 2215, 2301 CE Leiden, The Netherlands

SUMMARY

The Severity of Disabilities Scale (SDS) of the ICIDH reflects the degree to which an individual's ability to perform a certain activity is restricted. This paper describes the application of two models from item response theory (IRT), the graded response model and the partial credit model, in order to derive a tentative proposal for a revised SDS. The key ingredient of the approach is to scale existing disability items obtained in different studies on a common scale by exploiting the overlap. Both IRT models are fitted to a linked data set containing items for measuring walking disability. Based on these solutions, a tentative SDS is constructed. The paper concludes with a discussion of the implications, limitations and advantages of the approach. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

The International Classification of Impairments, Disabilities and Handicaps (ICIDH) [1] is a structured collection of personal and social consequences of functional limitations. The ICIDH is endorsed and promoted by the World Health Organization (WHO), and is used to classify persons and populations. A key component of the ICIDH Disability section is the Severity of Disabilities Scale (SDS). The SDS consists of a set of seven *severity codes*, and is meant to reflect the extent to which an individual's ability to perform a certain activity is restricted. Table I contains the current set of severity codes.

The psychometric properties of the SDS are not known, so one cannot say that the scale is unidimensional, that successive codes indicate equal steps in disability, or are even ordered. Transforming existing disability measurements into severity codes is problematic since no conversion keys are available. For these and a number of other reasons, [2, 3] the WHO commissioned TNO Prevention and Health to develop a revision of the SDS. The revised scale should increase in severity in approximately equal steps, it should enable the conversion of existing disability instruments into SDS codes, and it should provide a conversion rule from the current SDS.

*Correspondence to: S. van Buuren, TNO Prevention and Health, P.O. Box 2215, 2301 CE Leiden, The Netherlands.
†E-mail: S.vanBuuren@pg.tno.nl

Contract/grant sponsor: WHO Collaborating Centre

Received December 1998
Accepted May 2000

Table I. Current severity of disabilities scale of the ICIDH (Source: WHO [1]).

Code	Label	Includes
0	Not disabled	No disability present (the individual can perform the activity or sustain the behaviour unaided and on his own without difficulty)
1	Difficulty in performance	Difficulty present (the individual can perform the activity or sustain the behaviour unaided and on his own but only with difficulty)
2	Aided performance	Aid and appliance necessary (the individual can perform the activity only with a physical aid or appliance)
3	Assisted performance	The need for a helping hand (the individual can perform the activity or sustain the behaviour, whether augmented by aids or not, only with some assistance from another person)
4	Dependent performance	Complete dependence on the presence of another person (the individual can perform the activity or sustain the behaviour, but only when someone is with him most of the time). Excludes: inability
5	Augmented inability	Activity impossible to achieve other than with the help of another person, the latter needing an aid or appliance to enable him or her to provide this help (for example, the individual cannot get out of bed other than by the use of a hoist); behaviour can be sustained only in the presence of another person and in a protected environment
6	Complete inability	Activity or behaviour impossible to achieve or sustain (for example, an individual who is bed-bound is also unable to transfer)
8	Not applicable	
9	Severity unspecified	

Some innovations in educational statistics and psychometrics are directly relevant to this problem. Achieving comparability of scores on different forms of the same test is a classic problem in educational statistics. The problem is typically attacked by *test equating*, [4, 5] a statistical process that is used to adjust scores on test forms so that scores can be used interchangeably. Traditional equating methods derive conversion formulae for transforming one score into another. Such formulae are typically based on overlap in test forms and/or examinees. Modern forms of test equating rely on item response theory (IRT), a family of mathematical models for describing the properties of items rather than complete tests [6, 7]. Vale [8] and Baker [9] suggested that the equating problem can been formulated as a missing data problem in parameter estimation. One could, for example, collect both old and new data into one data set, and estimate item parameters from the combined data by disregarding those entries of items that are not administered. Software that can handle such estimation problems has become available only recently.

The present paper describes the application of IRT models to items for measuring walking disability. IRT models are helpful in placing the response categories of the items onto a common scale. Once the location of each category on the common scale is known, informed decisions can be made in light of the criteria set up for a revised SDS. The method assumes that items from different studies measure the same trait, that relevant microdata per study are available, and that each study has one or more items in common with other studies. Earlier work in this project is summarized in two technical reports [3, 10]. Related work in rheumatology and rehabilitation research has been done Tennant and McKenna [11] and Fisher *et al.* [12].

2. METHOD

The approach adopted to develop a proposal for a revised SDS consists of the following steps:

1. Make an inventory of items and instruments that intend to measure the same phenomenon, for example, walking disability.
2. Obtain data sets that contain observations on at least two items or instruments for measuring walking disability.
3. Combine the data sets into a linked data set.
4. Estimate the severity of disability pertaining to the response categories of each item on a common scale.
5. Order all estimates along the common scale, and classify item categories into homogeneous groups.

These steps are treated in more detail, in the following sections.

2.1. Instruments and items

Hopman-Rock and Miedema [3] identified 96 different instruments and items for measuring disability that are currently being used in areas such as population surveys, statistics, public health research, rehabilitation, vocational assessment and nursing homes. These instruments differed widely in quality, scope, popularity and coverage of the ICIDH. The authors identified a subgroup of 21 instruments that contained enough items relevant to the ICIDH disability codes 30 to 60. Items pertaining to these instruments were grouped into aspects according to the ICIDH disability classification. Disability categories that were measured most frequently were 'walking', 'dressing', 'disability in transfer to the toilet', 'bathing', 'other personal hygiene', 'feeding', 'climbing stairs', 'transfer' and 'subsistence'. The present paper is restricted to ICIDH disability code 40: 'walking'.

2.2. Data

Raw data at item level were obtained from various sources. The relevant literature was searched and the principal investigators of appropriate studies were contacted. Only studies that administered at least two of the 21 disability instruments were considered. In this way, data were obtained from Liang (*FIVE*) [13], Suurmeijer (*EURIDISS*) [14] and Hofman (*ERGOPLUS*) [15, 16]. We also used some of our own data (*GOW1*, *DETER*) [17, 18], and public microdata from the Netherlands Health Interview Survey 1994 (*CBS-GE*) [19]. Data were available on the following disability instruments: AIMS (Arthritis Impact Measurement Scale); FSI (Functional Independence Measure); SIP (Sickness Impact Profile); HAQ (Health Assessment Questionnaire); GARS (Groningen Activity Restriction Scale); ADL (Activities of Daily Living); OECD (OECD Long-Term Disabilities Questionnaire), and PPT (Physical Performance Test).

Table II contains the description of the items and the marginal frequency per category per study. Note that the sample sizes of studies *FIVE*, *GOW* and *DETER* are small. This has resulted in empty response categories for some items (for example, FSIH). Since information on the relative magnitude is lacking, these response categories cannot be placed onto a common scale.

Table II. Structure of the combined data set for walking disability based on six studies. Marginal counts per study are given. An empty block indicates that the specific item (in the row) was not administered in the study (in the column).

Item	Description	Categories	*FIVE* $n=38$	*ERGO+* $n=306$	*EURIDISS* $n=292$	*CBS-GE* $n=2113$	*GOW* $n=50$	*DETER* $n=30$
AIMS	Are you unable to walk unless you are assisted by another person or by a cane, etc...?	No	34					
		Yes	3					
FSIH	Walking inside	No help needed	34					
		Used a cane, special equipment or other device	1					
		Used someone's else's help	0					
		Unable to do the activity	0					
FSIP	Walking inside	No pain	20					
		Mild pain	7					
		Moderate pain	5					
		Severe pain	0					
		Extreme pain	0					
FSID	Walking inside	None	21					
		Mild difficulty	4					
		Moderate difficulty	4					
		Severe difficulty	0					
		Extreme difficulty	0					
SI01	I walk shorter distances or often stop for a rest	No	25	276				
		Yes	13	28				
SI07	I walk by myself but with some difficulty; for example I limp, wobble, stumble, ...	No	30	294				
		Yes	8	10				
SI08	I only walk with help from someone else	No	38	302				
		Yes	0	2				
SI11	I get about only by using a walking frame, crutches, stick, walls, or hold on to furniture	No	36	296				
		Yes	2	8				

SI12	I walk more slowly	No	20	244				
		Yes	18	60				
HAQ8	Are you able to walk outdoors on flat ground?	Without any difficulty	30	242	178			
		With some difficulty	7	43	68			
		With much difficulty	0	15	42			
		Unable to do	0	0	2			
GAR7	Can you, fully independently, get around in the house (if necessary, with a cane)?	I can do that by myself without any difficulty			207	1825		
		Idem.., with some difficulty			78	120		
		Idem.., with much difficulty			7	21		
		I can't, only with help of others			0	5		
GAR9	Can you, fully independently, walk outdoors (if necessary, with a cane)?	I can do that by myself without any difficulty			145	1606		
		Idem.., with some difficulty			110	229		
		Idem.., with much difficulty			29	86		
		I can't, only with help of others			8	52		
OECD	Are you able to walk 400 m without stopping? (if necessary with a cane)	Yes				1480	40	11
		Yes, with some difficulty				204	6	7
		Yes, with much difficulty				70	2	2
		No, I can't do that				169	2	10
PPT7	Walk 15 m	<=15 s					32	5
		15.5–20 s					14	6
		20.5–25 s					1	5
		>25 s					2	8
		Unable to do					1	5

2.3. Linkage structure

It is not informative to compare the responses on two items A and B if these items have been administered to different groups. Differences in the score distribution of A and B may be due to either differences between studies or to differences between items, or to a combination of both. However, if a third item C, that assesses the same trait, is measured in both studies, then the distribution of A and B can be compared through this common item. The idea is that differences between any two studies can be determined from items that are common to both.

As an example, consider item 1 of the SIP (SI01: I walk shorter distances or often stop for a rest) and item 9 of the GARS (GAR9: Can you, fully independently, walk outdoors (if necessary, with a cane)?) in Table II. The SI01 item has been administered in the *ERGOPLUS* study, the GAR9-item was collected in the *EURIDISS* study. Since both the samples and the items differ, there is no sensible way of comparing these distributions, but both studies also administered item number 8 of the HAQ (HAQ8: Are you able to walk outdoors on flat ground?). This item shows that the *EURIDISS* sample has more walking disabilities than the *ERGOPLUS* sample. The amount of severity that is measured by items SIO1 and GAR9 can now be compared through item HAQ8. It is said that SI01 and GAR9 are linked by HAQ8.

The organization of the data in Table II extends this principle to other studies and items. The table visualizes in what way walking items are distributed over different studies. An important use of this arrangement is to check if items are linked. Items are linked if there is a path connecting them. In educational testing, such a path is known as the *equating strain* (Kolen and Brennan, reference [5], p. 258). Table II displays a path from items AIM5 through PPT7, so these items are linked. In practice, one often needs to permute the rows and columns of the table in order to identify a connecting path.

Data of this structure could have been produced by a test equating design known as the *common-item nonequivalent groups design* [5] or the *anchor items design* [8]. These designs can provide information on differences between item and samples simultaneously. Unlike the situation in educational research however, the design is not under experimental control, and the exact form depends on the accidental overlap between studies. Thus, before gathering data and combining them into a common data set, it is necessary to investigate whether the items of interest can be linked. Additional data might be needed in some situations. In the worst case, one could be forced to plan a new study to sample the appropriate overlapping information.

2.4. Statistical models

Linked data contain relevant information for ordering both items and samples. This information is succinctly expressed as a set of parameters of a statistical model that is estimated from the data. This section describes two such models, both stemming from item response theory (IRT). The primary objective of these models is to explain the observed distribution of the responses as a function of a latent trait θ, here walking disability. This opens up the possibility to express severity of disability in terms of observable, empirical quantities.

Suppose that the set of possible answers on item j consists of m_j ordered categories, and let the response of person i on item j be represented by x_{ij}, which takes values $1, 2, \ldots, m_j$. Furthermore, let the ith respondent to characterized by an attribute value θ_i on a latent severity of disability dimension θ. Presumably, θ_i, influences x_{ij}. Two models that relate θ_i to x_{ij} are

considered: the graded response model [20] and the partial credit model [21]. The first model is somewhat easier to interpret and applies more naturally to our problem, while the second model has superior theoretical properties. We fit both models in order to get the best of both worlds.

The graded response model (GRM) assumes that, for a given item j, the probability of choosing a category k or higher (with $k=2,\ldots,m_j$) is specified as a logistic function of θ as

$$P(x_{ij}\geqslant k|\theta,a_j,b_{jk})=\frac{1}{1+\exp(-Da_j(\theta-b_{jk}))} \quad \text{for } k=2,\ldots,m_j$$

where a_j is a *slope parameter*, b_{jk} is a *category threshold parameter*, and D is a scale constant specifying the metric of the latent disability scale. The logistic curves of the same item are parallel and are allowed to vary in location only. In the following, we assume that $a_j=1$ for all items, primarily because the data that are to be analysed are sparse at some points. For a given item j, parameter b_{jk} can be interpreted as the θ-value at which exactly 50 per cent of the population scores in category k or higher. Computations are done in the conventional logistic metric where $D=1.7$. Let $P(x_{ij}\geqslant 1)=1$ and $P(x_{ij}\geqslant m_j+1)=0$. The probability of observing a specific category k for a given disability θ is then equal to

$$P(x_{ij}=k|\theta)=P(x_{ij}\geqslant k|\theta)-P(x_{ij}\geqslant k+1|\theta)$$

for all $k=1,\ldots,m_j$. These functions are referred to as *category characteristic curves.*

Let $P(x_{ij}=k-1|\theta)$ and $P(x_{ij}=k|\theta)$ denote the probability of observing category $k-1$ and k for a given disability θ. Given that the response occurs in either category $k-1$ or k, the partial credit model (PCM) defines the probability of choosing the kth category over the $k-1$th category as

$$\begin{aligned}P(x_{ij}=k|\theta)&=\frac{P(x_{ij}=k|\theta)}{P(x_{ij}=k|\theta)+P(x_{ij}=k-1|\theta)}\\&=\frac{\exp(Da_j(\theta-b_{jk}))}{1+\exp(Da_j(\theta-b_{jk}))}\end{aligned}$$

for $k=2,\ldots,m_j$. Rearranging the set of m_j-1 equations and defining $b_{j1}=0$ gives the probability of observing category k as a function of θ as

$$P(x_{ij}=k|\theta)=\frac{\exp(\sum_{c=1}^{k}Da_j(\theta-b_{jc}))}{\sum_{t=1}^{m_j}\exp(\sum_{c=1}^{t}Da_j(\theta-b_{jc}))}$$

for $k=1,\ldots,m_j$. The interpretation of a_j and D is similar to that in the GRM. The quantity b_{jk} is an *item step parameter* and gives the intersection point on θ for which the probabilities of obtaining responses $k-1$ and k are identical.

The following assumptions are needed to complete the model specification. First, for a fixed value of θ, the response probabilities are conditionally independent given θ, that is, $P(x_{ij}=k \text{ and } x_{ij^*}=k^*|\theta)=P(x_{ij}=k|\theta)P(x_{ij^*}=k^*|\theta)$ for $j\neq j^*$. Second, the latent trait θ is unidimensional, that is, items have only one dimension in common, as in a one factor model.

Table III. Parameter estimates of the graded response model (threshold parameters) and the partial credit model (item step parameters), ordered by value of the threshold parameter.

Walking item	Category	Graded response model MULTILOG	Graded response model PARSCALE	Partial credit model PARSCALE	Description of the upper category
GAR7	4	6.83	4.80	3.53	Inside: only with help
HAQ8	4	6.41	4.56	3.94	Outdoors: unable
SI08	2	5.41	3.78	3.73	Only with help
GAR7	3	4.72	3.51	2.85	Inside: much difficulty
GAR9	4	4.22	3.20	2.32	Outdoors: only with help
FSIH	2	3.96	2.97	2.87	Inside: used cane etc.
SI11	2	3.84	2.79	2.76	Use frame, crutches etc.
SI07	2	3.27	2.42	2.39	Limp, wobble, etc.
PPT7	5	2.91	2.33	1.80	Cannot walk 15 m
GAR9	3	2.87	2.26	1.85	Outdoors: much difficulty
HAQ8	3	2.73	2.21	1.76	Outdoors: much difficulty
FSID	3	2.74	2.19	1.49	Inside: moderate difficulty
FSIP	3	2.73	2.18	1.67	Inside: moderate pain
AIMS	2	2.67	2.14	2.07	Unable unless assisted
OECD	4	2.51	1.96	0.86	Cannot walk 400 m
GAR7	2	2.44	1.94	1.97	Inside: some difficulty
SI01	2	2.19	1.69	1.67	Shorter distances
OECD	3	2.03	1.61	1.44	400 m: much difficulty
PPT7	4	1.59	1.38	0.75	15 m: >25 s
FSID	2	1.38	1.28	1.49	Inside: mild difficulty
GAR9	2	1.29	1.10	1.24	Outside: some difficulty
SI12	2	1.28	1.05	1.05	More slowly
HAQ8	2	1.09	1.05	1.18	Outdoors: some difficulty
PPT7	3	1.06	0.99	1.26	15 m: 20–25 s
FSIP	2	0.93	0.97	1.11	Inside: mild pain
OECD	2	1.09	0.91	1.20	400 m: some difficulty
PPT7	2	−0.22	−0.03	0.23	15 m: 15–20 s

Third, statistical independence among respondents is assumed. The specifications of the GRM or the PCM can be used to define the likelihood of each observed response pattern in terms of θ and b_{jk}. The parameters of the GRM and the PCM are estimated by marginal maximum likelihood using PARSCALE 3.0 [22]. Missing values are coded as 'not presented' answers and are skipped during the computations. Latent disability estimates θ_i are derived by the expected *a posteriori* (EAP) estimator. The distribution is standardized to zero mean and unit variance.

3. ANALYSIS

Table III contains the estimated b_{jk}-parameters in both statistical models. The results are sorted by the threshold parameters of the GRM. PARSCALE converged in 16 iteration to a solution with -2 log-likelihood of 8376.204. As a check, MULTILOG 6.03 [23] was also used to estimate the parameters of the GRM. Apart from a linear transformation, the MULTILOG and PARSCALE solutions turned out to be virtually identical. The correlation between the

threshold parameters was equal to 0.998. The PCM solution took 17 iterations and produced a -2 log-likelihood equal to 8213.680. Although not apparent from Table III (because of different parameterizations), the results of the PCM are very similar to those of the GRM. The correlation between the θ-estimates of both models is equal to 0.994, so both models produce essentially identical disability estimates.

3.1. Interpretation of the model parameters

The top rows in Table III contain the threshold parameters for those category pairs that indicate serious disability. The threshold parameter is the level of disability at which 50 per cent of the people will respond in the higher category. Thus, for category 4 of the GAR7-item, we find a threshold of 4.8, which means that about half of a fictitious sample with an average disability of 4.8 (which is very high) will have a score in category 4. Category 4 of the GAR7 item measures more severe disability than the categories 4 of either the HAQ8, GAR9 or OECD items. Category pairs with high thresholds are usually associated with inability to walk, or walking only with the help of others. Indoor walking problems are considered more severe than the same amount of outdoor problems. In general, the ordering of response categories produced by the GRM seems to be sensible and interpretable, and matched our expectations.

Owing to different parameterizations, item step parameters as defined in the PCM cannot be directly compared to threshold parameters of the GRM if items have more than two categories. It is, however, possible to compare both models by their category characteristic curves. Figures 1 and 2 portray the response probability to each category separately as a function of disability θ. At every point along the θ-scale, the combined probability of all categories belonging to the same item adds up to 1. For example, for $\theta = 1$ the probabilities of observing categories 1, 2 and 3 of the FSIP-item are 0.47, 0.40 and 0.13, respectively. In the GRM, the maximal probability of category 2 (approximately 0.45) occurs at about $\theta = 1.6$. According to the PCM, the maximum is located at $\theta = 1.4$. Though the curves of the PCM are generally steeper and somewhat more reactive to θ, both models convey the same message.

Note that, especially in the GRM, answer categories 2 and 4 of the OECD are used more often than its third category for all levels of disability. Thus, according to the model, respondents prefer categories 2 or 4 to 3 irrespective their state of disability, which is not how a good response system should behave. If this were a test construction problem, one might try to alleviate the situation by, for example, combining categories 2 and 3 into a single category. In contrast to this, the characteristic curves of item GAR7 are approximately equally spaced over θ, while no answer category dominates another. Thus, the GAR7-item covers a large variation in disability. On the other hand, most curves of the PPT7-item, an item about how long it takes to walk 15 m, are located to the left side of the plot. This indicates that the item is sensitive to only relatively mild forms of disability.

3.2. Disability estimates

Figure 3 displays the distribution of walking disability per study on the common θ-scale. Distributions are generally quite skewed. In some cases (*ERGOPLUS*, *CBS-GE*, *GOW*), the median coincides with the 25th percentile. There are substantial differences in disability between the various sources. The order of mean walking disability of the studies from high to low was *DETER*, *FIVE*, *EURIDISS*, *CBS-GE*, *ERGOPLUS* and *GOW*. This ranking

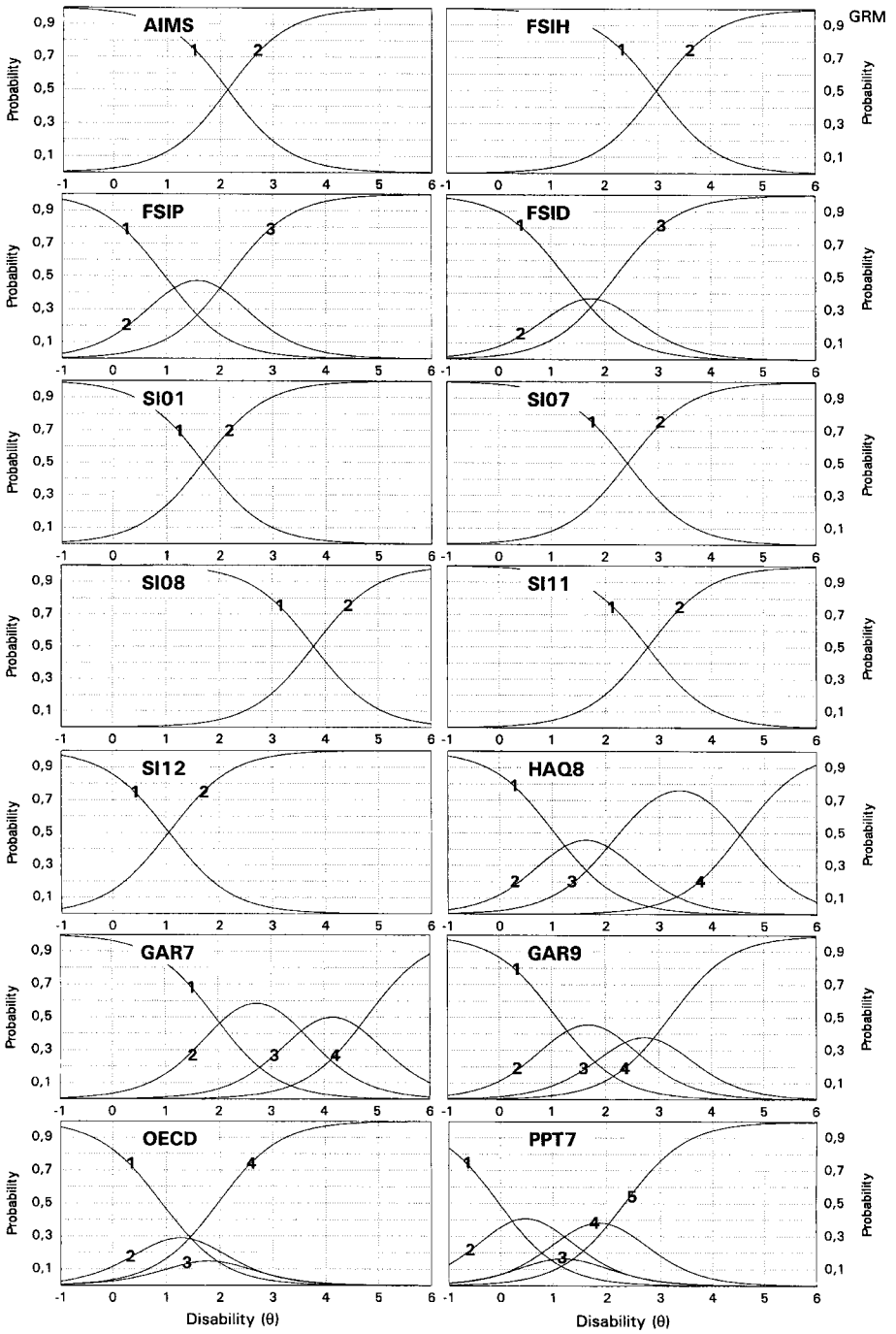

Figure 1. Category characteristic curves of the graded response model indicating the response probability for each item category as a function of disability θ.

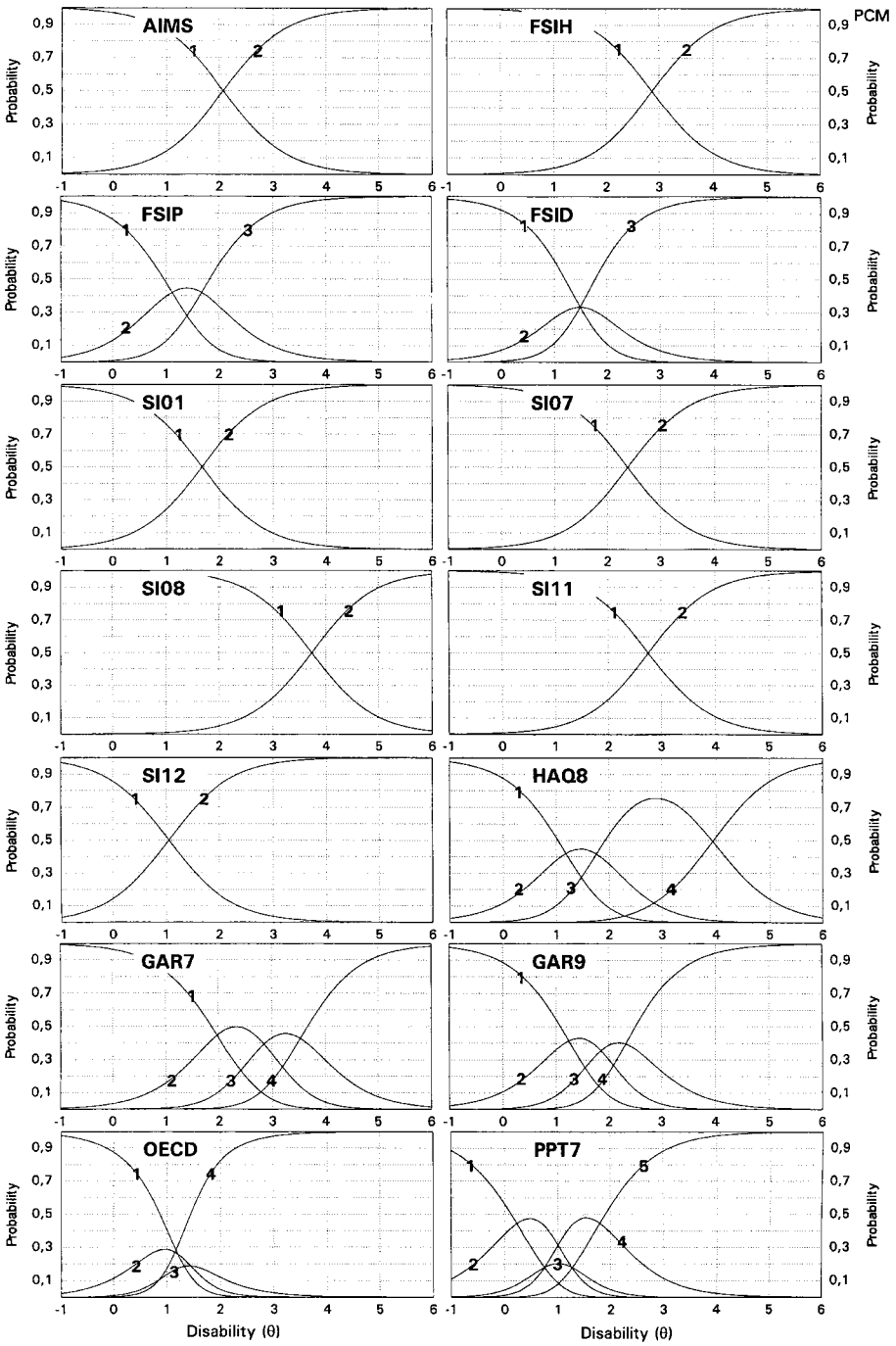

Figure 2. Category characteristic curves of the partial credit model indicating the response probability for each item category as a function of disability θ.

Table IV. Tentative proposal for the severity of disabilities scale of the ICIDH for walking disability.

Former code	Proposed code	Typical walking disabilities
0	0	No walking problems, person is able to walk 15 m in less than 15 s
1	1	Inside walking with mild pain, walking inside and outdoors with some or mild difficulty, walking more slowly, 400 m without stopping with some difficulty, 15 m in 15–25 s
1	2	Much difficulty walking outdoors, moderate difficulty walking inside, only short distances can be walked, cannot walk 400 m without stopping, 15 m in more than 25 s
2	3	Walking requires the use of an aid (cane, crutches, artificial limbs, walking frame etc.)
3,4	4	Walking outdoors is only possible with the help of someone else, and inside with much difficulty, cannot walk 15 m
4,5	5	Walking indoors is only possible with help of someone else, unable to walk outdoors
6	6	Completely unable to walk

agreed with our prior notions about the differences in ability between these samples. In fact, one of the authors predicted an almost identical ordering before seeing the actual results.

3.3. A tentative Severity of Disability Scale

Categories of different items are placed onto a common disability scale. This not only provides a means to compare responses across different items, but also suggests that a tentative ICIDH severity coding for walking disability can be found by dividing up the θ-axis into a number of groups. The wording of the most prominent response categories in a given θ-group may then act as empirical descriptive labels for the corresponding severity code.

Table IV provides a tentative severity classification into seven severity codes labelled 0–6. For each severity code, the table contains the proposed code, the corresponding current code, and a detailed description in terms of walking disability. Compared to the current coding, the proposed coding provides more room to discriminate mild levels of severity, while on the upper end of the scale, three disability codes are combined into one. It is somewhat unnatural to force a descriptive label as in Table I to each coding. An advantage of the proposed scheme is that the 'distances' in severity of disability are more even. Thus, one might say that the difference in disability between codes 1 and 2 is approximately equal to the difference between, say, codes 3 and 4. In other words, given the appropriateness of the IRT models the proposed scale is an interval scale.

4. DISCUSSION

A particular advantage of our approach is that only existing, though overlapping, data sets are needed. Explicit parameter estimates per category yield insight into the measurement characteristics of each item. The interpretation of the Severity of Disability Scale is facilitated as each severity code is defined in terms of observable empirical quantities. This opens up a way for measuring the severity of disability in a given person or population.

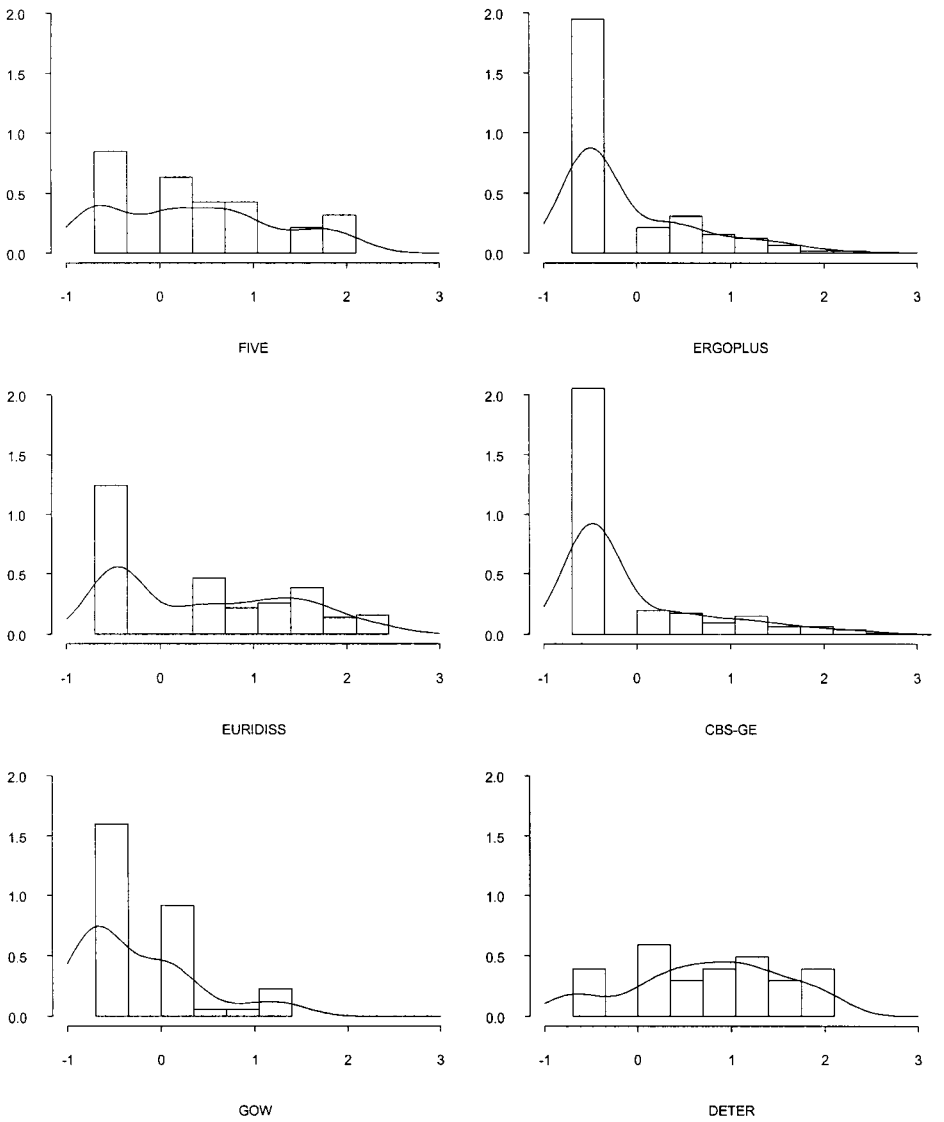

Figure 3. Frequency distribution of disability (θ) per study.

A limitation of the study is that it is based upon the analysis of only one form of disability – walking. Though our work on dressing disability [10, 24] confirmed the results for walking, the coverage of the ICIDH is still far from complete. A more extensive study into other aspects of disability would provide a better foundation for the proposal. Such a study should in particular deal with the question whether it is possible to derive a scale that is equally applicable to all forms of disability. We refer to Martin and Elliot [25] for a overview of the problems that one might encounter.

The samples contain few severely disabled people, and did not cover very disabled populations in, for instance, nursing homes or geriatric institutions. An often-touted virtue of IRT models is their sample independence, so one would expect similar parameter estimates for severely disabled samples. This applies especially to the PCM. However, since relatively few extremely disabled people were available in our samples, the upper end of θ is subject to random error. The accuracy of this upper end is likely to enhance if more severely disabled populations are included.

A controversial issue is whether severity of disabilities should be defined, measured and interpreted *with* or *without* aids and appliances. The analyses in this paper indicate that respondents usually considered performance with difficulty as less severe than performance with aids, which in turn is considered less severe than the need for a helping hand. This suggests that in practice severity of disabilities is more likely to be interpreted and measured as the severity without aids and appliances. We realize that the number of items on which this conclusion is based is small however.

The statistical analyses as done in this paper rely on a number of technical assumptions: unidimensionality; local stochastic independence; parallel item response functions; normality of the distribution of ability. It is not easy to check whether these conditions actually hold in our data. Several tests for the polytomous model have been developed [26], but most of these are still experimental and not available in the software we used. A complicating factor here is that our data have a linked structure, with a vast amount of missing data. It is not known if and how this affects the properties of the proposed tests.

Unidimensionality refers to the question whether items measure the same underlying trait. Since all items in the data evidently measure some aspect of walking disability, one could accept unidimensionality on face validity. Since such an approach is not satisfactory in general, a linear factor analysis on the incomplete correlation matrix was performed. Using the eigenvalue-larger-than-one criterion, two factors were identified. The first factor was clearly related to disability, but interpretation of the second was difficult. This analysis suggests that the set of items has a reasonable content validity, see van Buuren *et al.* [10] for details. PARSCALE prints a table of item fit statistics that could potentially be helpful in identifying badly fitting items. We noted however that these diagnostics depend very much on sample size, which diminishes their usefulness in linked data.

Another technical point is the normality of disability estimates. Though not mentioned in its manual, the estimation method used in PARSCALE assumes a normal ability distribution. Figure 3 shows that this is not true. No matter what the population is, disability distributions are nearly always skewed to the right. Other estimation methods that are insensitive to this distributional assumption have been proposed, but the question of just how robust the PARSCALE estimation method is against violations of the normality assumption is still debated.

A related issue is whether the ability distribution of each study group should be modelled separately. The rater's effect model (REM) was used to define an additional location shift parameter for each subgroup. The REM expresses the deviate of category k of item j in group g as $Z(\theta)_{gjk} = Da_j(\theta - b_{jk} - d_g)$, where d_g is the additional location shift parameter for the gth group. PARSCALE in 25 iterations ($-2\text{LL} = 7872.425$). Since the program could only estimate the parameters that are actually present in the specific subgroup, manual rescaling of d_g-parameters turned out to be necessary in order to properly place all subgroups on a common scale. The correlation between the disability estimates of the PCM and the REM

was equal to 0.98, which suggests that any bias associated with an improper choice of the disability distribution can only be small.

The combined data set contains a relatively long equating strain, which implies that the link of items on either end of the strain (AIMS, PPT7) is thin and fragile. This could prove especially troublesome when items are not unidimensional, as one aberrant item in the middle of the equating strain can effectively break the linkage. A solution is to use double linking [5]. Double linking designs have two or more paths from one item to another. Of course, it depends on the available data whether this is possible, but there is a potential gain in the validity of the scale.

Despite these limitations in data quality, scope and analyses, we feel that our method addresses an important topic in a proper way. A unique aspect of our approach is its empirical basis and its strong emphasis on items that are actually applied in the field. The interpretation and application of the proposed scale might therefore be easier than the current scale. The finer grain on the lower end of the proposed SDS enhances its suitability for applications in public health and prevention. The techniques used here provide a key to conversion issues. It is possible to translate the current SDS into the proposed SDS, to convert the severity as measured by existing disability items into the proposed SDS, and to convert existing items into other (existing or novel) disability items. Such possibilities will preserve much valuable work. Assets like these are likely to stimulate further developments regarding the scientific underpinnings of the ICIDH.

ACKNOWLEDGEMENTS

We thank the following people for their kind contribution of data: Dr M. H. Liang (Brigham and Women's Hospital, Boston), Dr T. P. B. M. Suurmeyer (Northern Centre for Health Care Reserch, Groningen), Dr E. Odding (Erasmus University, Rotterdam) and Dr C. Molleman (Higher Institute of Labour, Leuven). Data from the Netherlands Health Interview Survey (Netherlands Bureau of Statistics, Voorburg) were obtained through the Wetenschappelijk Statistisch Agentschap (WSA, Den Haag). We thank Marijke de Kleijn-de Vrankrijker, Harald Miedema and Jan van Rijckevorsel (TNO Prevention and Health, Leiden) for their stimulating role. Koos Zwinderman (Medical Statistics, Leiden University) and Peter van der Heijden (Methodology and Statistics, University of Utrecht) provided useful comments on a previous version of this work.

This project was finacially supported by a grant from the WHO Collaborating Centre for the ICIDH in the Netherlands.

REFERENCES

1. World Health Organisation (WHO). *International Classification of Impairments, Disabilities and Handicaps*. World Health Organisation: Geneva, 1980 (reprinted in 1993).
2. World Health Organisation and Statistics Netherlands (eds). *Third Consultation to Develop Common Methods and Instruments for Health Interview Surveys*. Statistics Netherlands: Voorburg, 1993.
3. Hopman-Rock M, Miedema HS. The development of a proposal for revision of the severity of disabilities scale of the ICIDH. Personal care, body disposition, locomotor and dexterity disbilities: Phase 1 and phase 2. Technical Report TNO/PG 95.043, TNO Prevention and Health, Leiden, 1995.
4. Holland PW, Rubin DB (eds). *Test Equating*. Academic Press: New York, 1982.
5. Kolen MJ, Brennan RL. *Test Equating: Methods and Practices*. Springer: New York, 1995.
6. Molenaar IW, Fischer GH (eds). *Rasch Models*. Springer: New York, 1995.
7. van der Linden WJ, Hambleton RK (eds). *Handbook of Modern Item Response Theory*. Springer: New York, 1996.
8. Vale CD. Linking items onto a common scale. *Applied Psychological Measurement* 1986; **10**:333–344.
9. Baker F. Equating tests under the graded response model. *Applied Psychological Measurement* 1992; **16**:87–96.

10. van Buuren S, Hopman-Rock M, Miedema HS. The development of a proposal for revision of the severity of disabilities scale of the ICIDH. Personal care, body disposition, locomotor and dexterity disabilities: Phase 3. Technical Report TNO/PG 96.024, TNO Prevention and Health, Leiden, 1996.
11. Tennant A, McKenna SP. Conceptualizing and defining outcome. *British Journal of Rheumatology* 1995; **34**:899–900.
12. Fisher Jr WP, Harvey RF, Taylor P, Kilgore KM, Kelly CK. Rehabits: a common language of functional assessment. *Archives of Physical and Medical Rehabilitation* 1995; **76**:113–122.
13. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Medical Care* 1990; **28**:632–642.
14. Suurmeijer TPBM, Doeglas DM, Moun T *et al*. The Groningen activity restriction scale for measuring disability: its utility in international comparisons. *American Journal of Public Health* 1994; **84**:1270–1273.
15. Hofman A, Grobbee DE, de Jong PTVM, van den Oudenland FA. Determinants of disease and disability in the elderly: the Rotterdam elderly study. *European Journal of Epidemiology* 1991; **7**:403–422.
16. Odding E, Valkenburg HA, Algra D, van den Oudenland FA, Grobbee DE, Hofman A. Association of locomotor complaints and disability in the Rotterdam study. *Annals of Rheumatic Diseases* 1995; **54**:721–725.
17. van Hell L, Hopman-Rock M. Ontwikkeling en evaluatie van het programma 'Goed Oud Worden': De testfase. Technical Report TNO/PG 95.040, TNO Prevention and Health, Leiden, 1995.
18. Hopman-Rock M. Determinanten van immobiliteit en fysieke activiteit: Een pilotstudie onder zelfstandig wonende ouderen die op de wachtlijst voor thuiszorg staan. Technical Report TNO/PG 94.005, TNO Prevention and Health, Leiden, 1994.
19. Statistics Netherlands. *Vademecum Gezondheidsstatistiek 1994*. Statistics Netherlands: Voorburg, 1995.
20. Samejima F. Estimation of latent ability using a response pattern of graded responses. *Psychometrika*, Monograph supplement, No. 17, 1969.
21. Masters GN. A Rasch model for partial credit scoring. *Psychometrika* 1982; **49**:529–544.
22. Muraki E, Bock RD. *PARSCALE*: *IRT based Test Scoring and Item Analysis for Graded Open-ended Exercises and Performance Tasks. Version 3*. Scientific Software International Inc.: Chicago, 1996.
23. Thissen D. *MULTILOG User's Guide. Version* 6.03. Scientific Software International Inc.: Chicago, 1991.
24. Hopman-Rock M, van Buuren S, de Kleijn-de Vrankrijker M. Polytomous Rasch analysis as a tool in the revision of the severity of disability scale of the ICIDH. *Disability and Rehabilitation* 2000; **22**:363–371.
25. Martin J, Elliot D. Creating an overall measure of severity of disability for the Office of Population Censuses and Survey Disability Survey. *Journal of the Royal Statistical Society*, *Series A* 1992; **155**:121–140.
26. Glas CAW, Verhelst ND. Tests of fit for polytomous Rasch models. In *Rasch Models*: *Foundations*, *Recent Developments and Applications*, Fischer CH, Molenaar IW, (eds). Springer, New York, pp. 325–352 (1995).

