Improved Interpretation of Stroke Trial Results Using Empirical Barthel Item Weights

Frank van Hartingsveld, MSc, PhT; Cees Lucas, PhD, PhT; Gert Kwakkel, PhD, PhT; Robert Lindeboom, PhD

Background and Purpose—Attempts have been made to provide guidelines for interpreting Barthel scores. We used a Rasch analysis to improve the measurement properties and clinical interpretability of the Barthel index score.

Methods—A specific extension of Rasch model was used to identify items that preclude the summation of items and to improve the item rating scale by examining the scores on the Barthel of 559 stroke patients scored 3 weeks (n=89) and 6 months (n=470) after stroke. The clinical interpretation of the revised Rasch modeled Barthel was illustrated by re-examining the results of a previously published trial on the effectiveness of leg and arm training after stroke.

Results—Most rating scales could be improved by collapsing nondiscriminating rating categories. Two items showed misfit: Bladder and Bowel. The remaining Barthel showed an excellent fit to the extended Rasch model (R1c Goodness-of-Fit P=0.35). Both items and patients could be placed on a common logit unit scale, allowing a clearer interpretation of the trial effect. Using the modeled activities of daily living difficulty/ability scale, we could express the differences between treatment arms in modeled probabilities of a positive score to each Barthel item for the treatment arms not conveyed by the original ordinal Barthel sum scores.

Conclusion—We improved the psychometric properties and clinical interpretation of the Barthel index. (Stroke. 2006;37: 162-166.)

Key Words: outcome assessment ■ psychometrics ■ stroke

The Barthel index was developed to monitor performance in activities of daily living (ADL) in stroke patients before and after treatment and to indicate the amount of nursing care needed. The Barthel has been used extensively as a measure of functional outcome in rehabilitation settings as well as in intervention studies.

As with most multi-item outcomes measures, the Barthel is based on an ordinal rather than an interval level scale. With ordinal scales, the overall score is obtained by simply adding up arbitrary numerical values assigned to a subject’s ratings on a series of items. For the Barthel,1 each item category is assumed equal and is scored using an arbitrary 5, 10, or 15 to arrive at a 0 to 100 scale or alternatively 0, 1, or 2 to arrive at a 0 to 20 scale. In practice, this implies that it is only possible to ascertain whether there has been a change in functional status. The exact amount of change, if any, cannot be determined, nor can it be interpreted in terms of functional ability.2 For this reason, attempts have been made to provide guidelines for interpreting Barthel scores.3

Modern scaling techniques, such as Rasch analysis, were specifically developed to undo the shortcomings of ordinal sum score instruments. Rasch analysis is based on a probability model to estimate “patient ability” and “item difficulty,” which are expressed on a common log-odds (logit) scale.4 This approach has some specific advantages over the traditional correlation-based psychometric techniques. First, it provides directions of how to weigh individual items to arrive at an interval level test score. Second, Rasch analysis incorporates strong statistical tests to identify items that do not belong to ADL domain and thus preclude the summation of items. Third, Rasch analysis provides directions how to combine item rating scale categories to obtain the best discrimination between the abilities of patients. Finally, the estimates of item difficulties and patient abilities are not affected by the sampled distributions of items and patients, and the Rasch-defined scale can be readily applied to new patient groups.5 Rasch analysis has been applied for the validation and modification of existing scales or for the validation of new scales in diverse areas of medicine including neurology.6

We used Rasch analysis on the Barthel scores of stroke patients: (1) to identify items that do not belong to the scale and thus preclude summation of the items, (2) to obtain empirical item weights for the Barthel, and (3) to determine the optimal rating scale for each item to maximize the sensitivity of the instrument to differences in ADL ability.
between patients. Our main objective was to address the limitation of a Barthel sum score in describing a patient in functional terms. We hypothesized that by placing patient and item measures on the same continuum, it should be possible to describe in functional terms a patient with a given sum score. Data of a randomized clinical trial (RCT) investigating the effects of arm and leg rehabilitation training on the functional recovery of stroke survivors was used to illustrate the interpretation of the revised scale.

### Materials and Methods

#### Patients

We used 2 data sets comprising the scores of 559 stroke patients on the Barthel from 2 previously published studies. The first data set consisted of 470 stroke patients who were scored on the Barthel 6 months after the event, on average 3 months after discharge from the hospital. They were the survivors of an original cohort of 760 consecutively admitted stroke patients who participated in a multicenter quality of care study in The Netherlands.\(^7\) The clinical and demographic characteristics of the sample are detailed previously. In short, mean age (SD) was 70.5 ± 12.5 years, 54% were male, 84% had an infarction, and 14% had a bleeding. The median (interquartile range [IQR]) Barthel score was 19 (15 to 20) points. The second data set was from an RCT investigating the effects of arm and leg rehabilitation training on the functional recovery of stroke inpatients using the Barthel as a primary end point. Mean age (SD) was 65 ± 10.6 years, 57.3% were male, and 100% had an infarction. The median (IQR) Barthel score was 7 (5 to 11) points.\(^8\) Patients (n = 89) were scored on the Barthel with weekly intervals; we used the scores of week 3 after stroke because there were no missing data on the Barthel at that occasion.

#### Statistical Analysis

An extension of the dichotomous 1-parameter logistic (Rasch) model for polytomous items\(^9\) was used to examine the scores of the subjects on the Barthel items. The dichotomous Rasch model can be used to describe the probability that subjects are scored in the positive direction in case of the Barthel to “can do task independently,” as a function of their ADL ability status. This probability is a function of the distance between the ability of subjects and the difficulty of the activity presented to them and is defined by the formula:

\[ P_i = \frac{e^{a_i(\theta - \beta_i)}}{1 + e^{a_i(\theta - \beta_i)}} \]

where \( \theta_i \) denotes the ADL ability level of patient \( k \), and \( a_i \) and \( \beta_i \) are known as the slope and the difficulty of the item \( i \), respectively. The extended Rasch model was used to identify items that do not belong to ADL scale and preclude the summation of items using \( \chi^2 \)-based goodness-of-fit-statistics.\(^9\) A fit statistic < 0.05 indicates item misfit. The overall fit of the remaining set of items to the model was examined using the \( R_{1c} \) statistic,\(^10\) where a \( P \) value > 0.05 indicates that the observed data have a satisfactory fit to the additive model.

#### Determination of Empirical Item Weights

The slope \( a_i \) of each Barthel item is imputed given the results of the fit statistic of an item and constitutes the empirical weight for a given Barthel item.\(^9\) After model fit has been established, the Rasch-weighted sum score is a sufficient statistic for the modeled ADL level \( \theta_i \) in logit units, which is expressed on the same scale as the items \( \beta_i \) and may range from −3 to +3. \( \theta_i \) and \( \beta_i \) were estimated using conditional likelihood estimation.\(^9\) The resulting hierarchy of items and patients aids in the interpretation of a weighted sum score. Using the probability model outlined above, one can predict the response pattern for a patient (eg, probability of “dressing independently”) from the weighted sum score.

#### Rating Scale Analysis

A particular focus in this study was to improve the Barthel item rating scales with >2 rating categories to maximize the sensitivity of the instrument to differences in ADL ability. To determine whether the rating scales were being used in a reliable manner, we examined the probability of each item score (eg, 0 to 3) in relation to the patients’ overall performance on the Barthel scale. This relationship was judged by plotting item category probability curves (“trace lines”). Disordered and thus unreliable rating categories were combined. Internal consistency reliability of the revised scale was evaluated with the Cronbach’s \( \alpha \)-coefficient. We hypothesized that the revised Barthel would have similar internal consistency reliability than the original scale, despite the possible removal of items and collapsing of rating scale categories.

### Results

#### Identification of Misfit Items

There were 2 misfitting items (\( \chi^2 \) Goodness-of-Fit; \( P < 0.05 \)): “Bowel” and “Bladder.” After removal of these items, weighting the remaining items and combining unreliable rating categories (see below), the items showed good fit to the extended Rasch model. The \( R_{1c} \) statistic \( P \) value was 0.35, indicating that the Rasch model was not rejected and that the items together form a 1D and additive scale.

#### Empirical Item Weights and Revised Item Rating Scales Based on Rasch Model Analysis

With the exception of the items “Feeding” and “Dressing,” all polytomous rating scales needed to be revised (Table 1). The highest item weight (\( a_i \)) was for the item “Toilet Use,” with a weight of 6. The item “Grooming” had the lowest weight (2). The weighted sum score of the revised Rasch unidimensional scale has a range of 0 to 45. The internal consistency coefficient for the revised 8-item Barthel was 0.93, which was identical to that of the original 10-item Barthel.

The Figure visualizes the logit unit item rating scale difficulties of the remaining Rasch homogeneous 8-item Barthel. Generally, the item difficulties were sufficiently spaced and measured different points on the underlying ADL ability construct. Only the “independent” category of the items “Mobility” and “Transfer” had practically the same difficulty (0.27 logits), meaning that they can be used interchangeably. The item difficulties in logit units ranged from −1.43 for the category threshold unable/needs help of the item “feeding” to 1.09 for the item “Stairs.”

#### Clinical Interpretation of the Revised Weighted Barthel Score

Table 2 displays for the 3 therapy arms of the “intensity of leg and arm training trial”\(^8\) the median sum scores at week 20 after stroke on the original Barthel (the primary end point), the median revised Rasch-weighted Barthel sum scores, and their associated median logit ADL ability measures. The original median score of the control group was 16, for the arm-training group 17, and for the leg-training group 19 points. The median revised Barthel-weighted score was 32.5 for the control group, 37.5 for the arm training, and 42.0 for the leg-training group. The median associated logit ADL ability measures were 0.494, 0.748, and 1.068, respectively (see Appendix for conversion table). The probabilities of passing an item rating category for each treatment arm are also presented. For the easiest items, the differences between the groups is only marginal. Almost all patients had a positive score on these items, irrespective of the therapy arm. For the
more difficult items, the differences in effectiveness between the therapy arms amounted to a >20% higher probability of a score in the positive direction for the leg training group compared with the arm-training group.

**Discussion**

Barthel data totaling 559 stroke patients showed good fit to the model. The overall fit of the revised scale was excellent (R1c statistic \( P = 0.35 \)), indicating that the extended Rasch model is a valid measurement model for the Barthel, and that the items together define a 1-dimensional ADL scale construct continuum. Two items, Bowel and Bladder, showed misfit. Misfit of the Bladder item in ADL scales was demonstrated previously.11,12 This is probably because all other items in Barthel refer to daily activities, whereas the items Bladder and Bowel refer to bodily functions ("impairments"). However, Hsueh et al13 did not find misfit for the Bladder items in the Barthel, possibly because of the smaller sample size \((n = 245)\) resulting in a lack of power to reject the measurement model. The maintenance of misfit items such as Bladder and Bowel, although important from a clinical or prognostic perspective,14 may lead to distorted results when used for the assessment of the ADL level of patients. Therefore, perhaps it is better to use bladder and bowel functioning as a separate clinical end point in the evaluation of therapy effectiveness. For example, one could report the difference in the rates of incontinence between the trial arms being compared.

Rasch analysis greatly improved the interpretation of test scores. Items and patients were placed on a common logit unit scale, allowing a clearer interpretation of trial effects (Figure; Appendix). The Rasch-weighted test score–associated logit measures can be converted into probabilities of performing certain ADL tasks. Using the probabilities, the clinical meaning of a score or score improvement can be judged in a more straightforward manner. Using the example of a published RCT,8 a score difference of 5 points, from 37 to 42 on the revised Barthel between the therapy arms (Table 2),
Another objective was to use Rasch analysis to determine whether the item rating categories were used in the expected manner. Inspection of the category probability curves showed that this was not the case for the majority of the Barthel items (Table 1). Some rating categories were underused, possibly because of the vague category descriptors (eg, “minor help” and “major help” for the item Transfer. Also, some item rating categories do not serve as a useful point of differentiation. For example, for the item Mobility, the adjacent rating categories “Wheelchair” and “With Help” may not be unambiguous, since many patients use a wheelchair while able to walk short distances. The resulting lack of reliability in the rating categories compromise the ability of the scale to discriminate among the ADL ability of patients. Most Barthel item rating scales needed to be revised. The internal consistency reliability (Cronbach’s α) of the resulting revised 8-item scale was unaltered (0.93) despite the collapse of several item rating scale categories and the removal of the 2 items, which supports the validity of the Rasch analysis. The improved scoring, with a 1-on-1 relationship with the logit unit ADL disability measure, may lead to smaller sample sizes needed to detect therapy effects. As a proof of principle, the reader may, using the revised scoring, find out whether this is indeed true.

A limitation in our study was that we used a Dutch sample. Further study is necessary to confirm the utility of the item weights and stability item difficulty calibrations in eg, English-speaking patients. However, previous Rasch analysis of the Barthel in a UK sample and Taiwanese sample showed an item difficulty hierarchy that was remarkably similar to ours. Because the Barthel is performance based, we do not expect that the item statistics would be substantially different in non-Dutch samples. In the meantime, the revised Barthel should be used to measure ADL functioning in new people with stroke. The scale in its present form may be used as a sensitive and linear outcome measure in stroke clinical trials or to monitor progress after stroke.

### Appendix

An extended Rasch-model weighted Barthel sum score $S$ can be computed by multiplying each Barthel item score with its discrimination parameter $a_i$ (Table 2) and sum the results:

$$S = \sum a_i * \text{item}$$

The weighted Barthel score $S$ is a sufficient statistic for a patient’s ADL ability measure $\theta$ in logit units (ie, the score that is most likely to occur given a patient’s latent [not directly observable or measurable] ADL ability level $\theta$ [Table 3], which is expressed on the same logit scale as the item difficulties).

### TABLE 2. Differences in Efficacy Between the Study Arms of the “Intensity of Arm and Leg Rehabilitation Training” Study (8)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Control (n=34)</th>
<th>Arm Training (n=29)</th>
<th>Leg Training (n=26)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median raw Barthel score (range 0–20)</td>
<td>16</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>Median revised Barthel-weighted score (range 0–45)</td>
<td>32.5</td>
<td>37.5</td>
<td>42.0</td>
</tr>
<tr>
<td>Median logit measure</td>
<td>0.404</td>
<td>0.748</td>
<td>1.068</td>
</tr>
</tbody>
</table>

*To calculate the probability of a given item category score for the “median” patient in each arm, the equation of the extended Rasch model with the concerning values of $\theta$, $b_i$, and $a_i$ was used (see Appendix and Figure).

### TABLE 3. Weighted Score $S$ and Interval Scale ADL Ability Measure $\theta$

<table>
<thead>
<tr>
<th>$S$</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>28</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>33</td>
</tr>
<tr>
<td>11</td>
<td>34</td>
</tr>
<tr>
<td>12</td>
<td>35</td>
</tr>
<tr>
<td>13</td>
<td>36</td>
</tr>
<tr>
<td>14</td>
<td>37</td>
</tr>
<tr>
<td>15</td>
<td>38</td>
</tr>
<tr>
<td>16</td>
<td>39</td>
</tr>
<tr>
<td>17</td>
<td>40</td>
</tr>
<tr>
<td>18</td>
<td>41</td>
</tr>
<tr>
<td>19</td>
<td>42</td>
</tr>
<tr>
<td>20</td>
<td>43</td>
</tr>
<tr>
<td>21</td>
<td>44</td>
</tr>
<tr>
<td>22</td>
<td>45</td>
</tr>
</tbody>
</table>
Acknowledgments
We thank Dr Rob J. de Haan for his kind permission for using his data. The authors would like to thank Drs. Martin van der Esch, Raymond Ostelo, and Jos C.M. van de Nes for their critical comments on earlier drafts of the manuscript.

References
Improved Interpretation of Stroke Trial Results Using Empirical Barthel Item Weights
Frank van Hartingsveld, Cees Lucas, Gert Kwakkel and Robert Lindeboom

Stroke. 2006;37:162-166; originally published online November 23, 2005;
doi: 10.1161/01.STR.0000195176.50830.b6
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2005 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://stroke.ahajournals.org/content/37/1/162

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org/subscriptions/